Análise de redes sociais: Aplicação a uma rede de clientes

por

Américo José Caulino Guerreiro

Tese de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Orientada por

Professor Doutor João Manuel Portela da Gama

Faculdade de Economia

Universidade do Porto

2012

À minha mulher Joana e ao pequeno Manuel

Agradecimentos

Gostaria de agradecer ao Professor Doutor João Gama pela orientação, disponibilidade e paciência para responder a todas as minhas questões.

No meu emprego, um apreço especial à minha chefe, Clara Carvalho, pelo apoio e compreensão, principalmente nesta última fase do meu trabalho. Agradeço também à minha colega Ernestina pela ajuda na forma de ideias trocadas em várias conversas.

A toda a minha família pela compreensão durante as minhas ausências e incentivo durante as minhas presenças.

Finalmente gostaria de agradecer à minha mulher Joana por todo o carinho, compreensão e apoio incondicional em todos os passos desta caminhada.

A todos um Muito Obrigado.

Resumo

Neste trabalho é descrita uma metodologia de análise de redes de clientes na óptica de redes sociais. Com esta abordagem é possível fazer análises a dados de clientes de uma forma completamente diferente dos métodos tradicionais (mesmo dos utilizados em análise de dados).

É proposta uma metodologia e métricas específicas de análise de redes sociais bem como a interpretação destas métricas à luz de uma rede de clientes. São propostas métricas de análise ao nível dos nós, da rede e é também feita a detecção de comunidades.

Para além das medidas propostas é feita uma análise dinâmica da rede. A abordagem da dinâmica temporal é composta por duas perspectivas: janela temporal deslizante e janela temporal acumulada. Estes dois métodos permitem analisar a dinâmica da rede de forma distinta. Enquanto uma tem em consideração os registos passados, a outra introduz uma componente de envelhecimento dando mais importância às transformações mais recentes na rede.

O método é finalmente aplicado a uma base de dados de CRM (*Customer Relationship Management*) de uma empresa.

Palavras-Chave: Análise Redes Sociais, Análise Redes Clientes, Análise Temporal Redes, Janela Acumulada, Janela Deslizante, CRM, Redes Dinâmicas

Abstract

This paper comprises the description of a methodology for the analysis of Customer Networks in the scope of Social Networks Analysis. With this approach it is possible to perform data analysis to customer databases in a completely different way from the traditional methods (even from those used in data mining).

Several metrics and measures used in Social Networks analysis are purposed and interpreted in the light of a Customer Network. The method uses metrics on the node and network level and community detection algorithms.

Beyond the purposed metrics, a dynamic analysis is performed. The time-driven dynamic is performed in two distinct ways: sliding window and cumulative window. These two approaches allow different analysis. While one takes past records into consideration, the other introduces an aging component associating a greater importance to more recent events.

The method is finally applied to the CRM (Customer Relationship Management) database of a real company.

Keywords: Social Network Analysis, Customer Network Analysis, Network Evolution Analysis, Sliding Window, Cumulative Window, CRM, Dynamic Networks

Índice

Capítulo 1 1
1 Introdução 1
1.1 Motivação 1
1.2 Objectivos 2
1.3 Contribuições 3
1.4 Organização 3
Capítulo 2 4
2 Análise de Redes 4
2.1 História e conceitos gerais 4
2.2 Tipos de Redes 6
2.3 Medidas estatísticas 7
2.3.1 Medidas ao nível dos nós 8
2.3.2 Medidas ao nível da rede 10
2.4 Modelos de Redes 13
2.4.1 Grafo aleatório 13
2.4.2 Modelo “pequeno mundo” 14
2.5 Propriedades das Redes 14
2.6 Comunidades 19
2.6.1 Detecção de comunidades 20
2.7 Análise de redes dinâmicas 24
2.7.1 Introdução às redes dinâmicas 24
2.7.2 Detecção de comunidades em redes dinâmicas 24
Capítulo 3 26
3 Metodologia 26
3.1 Representação da Rede 26
3.2 Análise da Rede 26
3.2.1 Métricas ao nível dos nós 27
3.2.2 Métricas ao nível da rede 29
3.2.3 Detecção de comunidades 30
3.2.4 Dinâmica temporal 30
3.2.5 Dinâmica temporal e detecção de comunidades 32

Capítulo 4 35
4 Estudo de Caso 35
4.1 Análise da rede 35
4.2 Análise janela deslizante 35
4.2.1 Apresentação visual 35
4.2.2 Medidas ao nível dos nós 37
4.2.3 Medidas ao nível da rede 41
4.2.4 Detecção comunidades 42
4.3 Análise com janela acumulada 46
4.3.1 Apresentação visual 46
4.3.2 Medidas ao nível dos nós 48
4.3.3 Medidas ao nível da rede 52
4.3.4 Detecção comunidades 53
4.4 Análise comparativa dos resultados 58
4.5 Sumário das duas abordagens 60
Capítulo 5 61
5 Conclusões 61
5.1 Lições aprendidas 61
5.2 Trabalho Futuro 62
Referências Bibliográficas 63

Índice de figuras

Figura 2.1: Exemplo rede criada por Euler 4
Figura 2.2: Exemplo de um grafo 6
Figura 2.3: Matrizes de Adjacência, Incidência e tabela de Adjacências 6
Figura 2.4: Exemplo de dois *random graphs* p=0.04 e p=1 13
Figura 2.5: Rede "pequeno mundo" (modelo Watz-Strogatz) 14
Figura 2.6: Aplicação algoritmo Blondel. Esquerda: Rede inicial; Centro: Cálculo de modularidade; Direita: Agrupamento comunidades 23
Figura 3.1: Metodologia proposta 27
Figura 3.2: Janela deslizante (esquerda); janela acumulada (direita) 31
Figura 3.3: Evolução comunidades 33
Figura 4.1: Janela deslizante 36
Figura 4.2: Grau e intermediação - P00003905 38
Figura 4.3: Grau e intermediação - P00000764 38
Figura 4.4: Grau e intermediação - P00002460 39
Figura 4.5: Grau e intermediação - X00000002 39
Figura 4.6: Grau e intermediação - P00015123 39
Figura 4.7: Grau e intermediação - P00009258 40
Figura 4.8: Centralidade *eigenvector* 41
Figura 4.9: Representação da rede com detecção de comunidades com janela deslizante 43
Figura 4.10: Representação da rede com primeiro agrupamento de comunidade com janela deslizante 44
Figura 4.11: Evolução das comunidades com janela deslizante 45
Figura 4.12: Janela acumulada. Instantes 1-3 a 6-8. 46
Figura 4.13: Janela acumulada. Instantes 7-9 a 10-12. 47
Figura 4.14: Top 5 de Grau e Intermediação - P00003905 48
Figura 4.15: Top 5 de Grau e Intermediação - P00001754 49
Figura 4.16: Top 5 de Grau e Intermediação - P00004323 49
Figura 4.17: Top 5 de Grau e Intermediação - P00007332 49
Figura 4.18: Top 5 de Grau e Intermediação - P00028619 50
Figura 4.19: Grau e intermediação - X00000002 50

Figura 4.20: Grau e intermediação - P00015123 51

Figura 4.21: Grau e intermediação - P00009258 51

Figura 4.22: Centralidade *eigenvector* 52

Figura 4.23: Representação da rede com detecção de comunidades com janela acumulada – instantes 1-3 a 1-11 54

Figura 4.24: Representação da rede com detecção de comunidades com janela acumulada – instante 1-12 55

Figura 4.25: Representação da rede com primeiro agrupamento de comunidade com janela acumulada – instantes 1-3 a 1-8 55

Figura 4.26: Representação da rede com primeiro agrupamento de comunidade com janela acumulada – instantes 1-9 a 1-12 56

Figura 4.27: Evolução das comunidades com janela acumulada..... 57

Capítulo 1

1 Introdução

Neste capítulo pretende-se dar uma visão geral dos motivos e objectivos pretendidos com a realização desta tese de dissertação. Serão explicadas as contribuições deste trabalho para a temática da análise de redes e a estrutura e organização deste documento.

1.1 Motivação

Vivemos numa aldeia global. Hoje, mais do que em qualquer altura da nossa história, as fronteiras e distâncias são mínimas. A velocidade a que a tecnologia tem evoluído permitiu à sociedade viver, comunicar e trabalhar de uma forma global. Com a abertura que existe nos mercados, hoje é relativamente fácil para uma empresa iniciar uma actividade em qualquer parte do mundo (partindo do princípio que tem os fundos necessários). Apesar das vantagens associadas a esta conjuntura, também existem desvantagens. A concorrência entre empresas é cada vez maior e, na competição para conquistar clientes, qualquer ponto adicional poderá fazer a diferença entre conseguir ganhar ou perder um negócio.

O conhecimento do comportamento ou características de um cliente faz, naturalmente, pesar a balança para o lado positivo podendo ser o factor influenciador para fazer um negócio.

Com as aplicações de CRM (Customer Relationship Management) disponíveis actualmente, as empresas tem a possibilidade de guardar bastante informação sobre os seus clientes tais como localização, hábitos de compra de produtos, valores de propostas, etc. Com o auxílio de técnicas de *Data Mining* estes dados podem ser

trabalhados de forma a conseguir obter informação que não é possível extrair dos dados através de análises tradicionais.

Neste trabalho propõe-se fazer a análise de redes de clientes aplicando técnicas de análise de redes sociais. Desta forma e através de detecção de características da rede, podem-se revelar alguns comportamentos ou tendências semelhantes aos observados em outros tipos de redes que permitam tirar conclusões que possibilitem o aumento da vantagem competitiva de uma empresa.

A proposta para a criação da rede é baseada nas premissas em que um cliente é um nó e um produto ou linha de produto é uma ligação.

1.2 Objectivos

Actualmente existem várias formas e métodos para análise de dados de clientes. Para além dos métodos tradicionais como análises estatísticas simples, estão também disponíveis técnicas próprias de análise de dados como análise de agrupamentos. Na pesquisa efectuada para a execução deste trabalho não foram encontradas referências para a análise de dados de clientes na óptica de redes sociais em que as ligações entre os clientes são os produtos adquiridos. O objectivo deste trabalho é desenvolver uma metodologia que permita analisar redes de clientes como uma rede social, tirando partido das técnicas de análise disponíveis e aplicáveis a esta área. Para além do conceito de redes sociais, a metodologia engloba duas perspectivas diferentes de análise temporal, uma com a aplicação de janela deslizante e outra com a aplicação de janela acumulada.

1.3 Contribuições

A principal contribuição desta tese é uma metodologia de análise de redes de dados de clientes na óptica de redes sociais englobando duas formas complementares de análise temporal.

No que diz respeito à análise de rede, são utilizadas métricas que permitem obter informação dos próprios nós e da rede em si. Com a interpretação das métricas é possível tirar conclusões relativas aos clientes e ao seu comportamento.

A detecção de comunidades e a dinâmica temporal complementam a análise permitindo visualizar ao longo do tempo a evolução da rede e das comunidades de clientes. Adicionalmente, pela aplicação de duas técnicas temporais distintas, a análise temporal permite uma análise ainda mais completa.

1.4 Organização

Esta dissertação encontra-se dividida em quatro partes. No Capítulo Dois é feita um levantamento do "estado da arte" relativamente aos temas de Redes e Redes Sociais, no Capítulo Três é apresentada uma metodologia para a aplicação a uma rede de clientes à luz dos conceitos revistos, no Capítulo Quatro é aplicada a metodologia proposta numa rede real e no Capítulo Cinco são apresentadas as conclusões e perspectivas de trabalho futuros.

Capítulo 2

2 Análise de Redes

A generalização da utilização de redes sociais como por exemplo, o Facebook ou Hi5 tem dado uma crescente relevância à análise de redes e mais especificamente à análise de redes sociais. Existem, no entanto, redes de outros tipos e natureza diferente das redes sociais sobre as quais é possível realizar as mesmas análises. Um exemplo e motivação para a realização deste trabalho são redes de clientes.

2.1 História e conceitos gerais

O conceito de rede é um termo bastante antigo. Uma das suas primeiras referências data de 1766 (Euler, 1766). Leonhard Euler tentou resolver o problema da travessia da cidade de Königsberg através das suas sete pontes desenhando uma rede de ligações. De forma a conseguir analisar os vários percursos possíveis representou os vários pontos da cidade ligados por linhas (Figura 2.1).

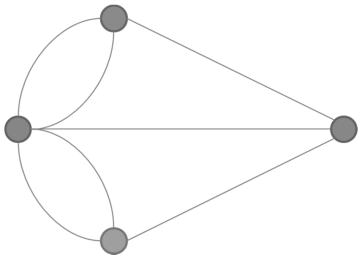

Figura 2.1: Exemplo rede criada por Euler

A primeira referência para o conceito de grafo data de 1878 (Sylvester, 1878) sendo que a primeira publicação sobre a teoria dos grafos surgiu em 1936 do autor Dénes Kőnig (Tutte, 2001).

Segundo a teoria dos grafos, um grafo é composto por um conjunto de objectos ligados entre si. Os objectos são denominados de vértices enquanto que as ligações, caracterizando uma relação entre os objectos, são denominadas de arestas. Para além das denominações vértice-aresta usadas na caracterização do sistema também é possível encontrar as designações equivalentes nó-ligação (Albert & Barabási, 2002) (Easley & Kleinberg, 2010). A representação formal de um grafo G é G=(V(G),A(G)), com conjuntos não vazios de V(G) vértices e A(G) arestas (Gama & Oliveira, 2010). As ligações entre os grafos podem ser direccionadas ou não direccionadas e dependem naturalmente do modelo que está a ser representado. Exemplos de grafos direccionados são ligações aéreas entre cidades, padrões de chamadas telefónicas e diagramas de estados (Berry, Linoff, 1997). Associado às ligações também pode haver um peso, $w \in \mathbf{R}^+_0$, sendo que este permite caracterizar a força da ligação. Desta forma os grafos são designados de pesados ou não pesados (Gama & Oliveira, 2010).

Um grafo G pode ser representado por uma matriz de adjacência A_{ij}, quadrada, com n^2 elementos, em que $A_{ij}=A_{ji}$ e i e j representam nós do grafo. O valor de cada A_{ij} pode ser 0, no caso de não haver ligação entre os nós i e j, 1 para uma ligação entre i e j e qualquer $w \in \mathbf{R}^+_0$ para o caso de grafos pesados. A diagonal desta matriz simboliza a ligação de nós com eles próprios, denominados de laços. Esta matriz é simétrica no caso de o grafo ser não-direccionado. Outra representação para um grafo é feita através de uma matriz de incidência I_{kl} em que k e l representam os nós e as ligações respectivamente. Nesta matriz é representada a incidência de um nó numa aresta. O valor de I_{kl} também depende de ser um grafo pesado e assume o valor 0, no caso de não haver incidência entre o nó k e a aresta l e 1 para uma ligação entre i e j. Para o caso de grafos pesados assume qualquer valor $w \in \mathbf{R}^+_0$ (Gama & Oliveira, 2010) (Newman, 2003).

Para grafos direccionados também é apresentada uma lista de adjacências indicando a direcção das ligações (Gama & Oliveira, 2010). Pode-se observar um exemplo de um grafo e respectivas matrizes na Figura 2.2 e Figura 2.3.

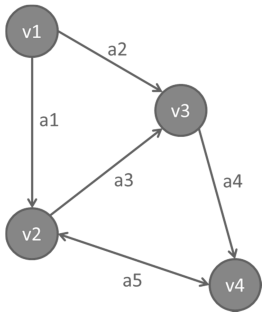

Figura 2.2: Exemplo de um grafo

	v1	v2	v3	v4
v1	0	1	1	0
v2	1	0	1	0
v3	1	1	0	1
v4	0	0	1	0

	a1	a2	a3	a4
v1	1	1	0	0
v2	1	0	1	0
v3	0	1	1	1
v4	0	0	0	1

vértice	vértice
v1	v3
v1	v2
v2	v3
v2	v5
v3	v4
v4	v2

Figura 2.3: Matrizes de Adjacência, Incidência e tabela de Adjacências

2.2 Tipos de Redes

Existem vários tipos de estruturas presentes no nosso quotidiano que podem ser analisadas como uma rede. Actualmente, com a proliferação de redes sociais como o Facebook, Tweeter, Flickr, Google+ entre outros, estes conceitos são bastante aplicados à análise do comportamento social. Existem, no entanto, outros tipos de estruturas reais que podem ser consideradas redes. De acordo com Newman, as redes reais podem ser

categorizadas em quatro tipos: **redes sociais**, **redes de informação**, **redes tecnológicas** ou **redes biológicas**.

Uma **rede social** é um conjunto de pessoas ou grupos de pessoas com algum tipo de padrão de contacto ou interacção entre eles. Nesta classificação enquadram-se, por exemplo, redes de amizade, redes de negócios entre empresas, redes de comunicação informal dentro de empresas (Gama & Oliveira, 2010) (Newman, 2003).

Quanto a **redes de informação** ou também denominadas de redes de conhecimento, tem-se o exemplo clássico da rede de citações entre documentos académicos. Nestas redes os nós são os documentos académicos e as ligações são a referência que um documento faz ao outro. Uma característica particular deste tipo de rede é o facto de ser acíclica uma vez que um documento só poderá fazer referência a documentos passados e não a documentos futuros. Outro exemplo é o caso da própria World Wide Web. Neste caso, como existe a possibilidade de referências cruzadas entre páginas, o modelo é cíclico (Gama & Oliveira, 2010) (Newman, 2003).

As **redes tecnológicas** são criadas pelo homem, construídas para a distribuição de recursos, mercadorias ou produtos de consumo. Exemplos deste tipo de redes são as redes eléctricas, redes de computadores, ferroviárias, etc. (Gama & Oliveira, 2010) (Newman, 2003).

A última classificação referenciada por Newman, **redes biológicas**, é referente a processos biológicos, tal como relações metabólicas dentro de uma célula, interacções físicas entre proteínas. Nesta classe também se enquadram redes neuronais (Gama & Oliveira, 2010) (Newman, 2003).

2.3 Medidas estatísticas

O estudo ou análise que é feita a uma rede tem como principal objectivo identificar e compreender o comportamento do sistema que originou a rede. Este estudo é muitas vezes efectuado recorrendo a métricas e medidas originárias do campo da Estatística.

Estas medidas são bastante úteis uma vez que permitem obter informação da rede sem a necessidade de saber a sua representação gráfica.

As medidas podem ser divididas de acordo com o tipo de análise que se pretende efectuar, isto é, se a análise é feita ao nível de um nó, permitindo descobrir a sua importância na rede em geral ou, se a análise é feita ao nível da própria rede, identificando características da estrutura global da rede e do comportamento que a gerou (Gama & Oliveira, 2010).

2.3.1 Medidas ao nível dos nós

Tal como explicado anteriormente, este tipo de medidas permitem identificar a importância de um nó dentro da estrutura da rede. Desta forma consegue-se obter quais os elementos mais importantes ou mais influentes na rede. Sendo assim, o primeiro objectivo de uma análise a este nível será percepcionar a **centralidade** ou **prestígio** de um nó. As métricas mais utilizadas para este tipo de análise são **grau**, **intermediação**, **proximidade** e **centralidade *eigenvector*** (Oliveira & Gama, 2010).

Grau é obtido pelo número de ligações num dado nó. Esta medida tem duas variantes: **grau interior** designado por k_i^+ e **grau exterior** designado por k_i^-**. Grau interior** é o número obtido pelo número de ligações com início no nó *n* enquanto que **grau exterior** é o número de ligações que terminam no nó *n*. As fórmulas destas duas métricas estão representadas em baixo.

$$k_i^+ = \sum_{j=1}^{n} a_{ji} \qquad (2.1)$$

$$k_i^- = \sum_{j=1}^{n} a_{ij} \qquad (2.2)$$

Sendo que o valor de k varia entre zero e n.

Na análise de redes sociais, esta métrica assume a designação de prestígio em que **grau interior** é denominado de suporte e **grau exterior** de influência (Freeman, 1979) (Oliveira & Gama, 2010).

Esta métrica é apenas aplicável a redes **não pesadas**. A utilização em redes **pesadas** é denominada de **força** e é dada pela fórmula 2.3.

$$k_i^w = \sum_{j=1}^{n} a^w{}_{ij} \qquad (2.3)$$

Intermediação permite medir o quanto um determinado nó está entre outros nós da rede. Esta medida permite inferir se um nó ocupa uma posição crítica na rede, ou seja, se a comunicação entre diferentes grupos passa por este nó. Um nó com um grau elevado de **intermediação** ocupa uma posição chave numa rede e é normalmente denominado de *gatekeeper*. A fórmula do cálculo de **intermediação** é dada pela seguinte equação:

$$\boldsymbol{b^v} = \sum_{\boldsymbol{s,t} \in \boldsymbol{V(G)\setminus v}} \frac{\boldsymbol{\sigma_{st}(v)}}{\boldsymbol{\sigma_{st}}} \qquad (2.4)$$

em que $\sigma_{st}(v)$ é o número de caminhos mais curtos que passam pelo nó *v* enquanto que $\boldsymbol{\sigma_{st}}$ é o número de caminhos mais curtos que passam pelos nós *s* e *t* (Oliveira & Gama, 2010).

Esta métrica também pode ser aplicada a ligações sendo que a **intermediação** de uma ligação é definida pelo número de caminhos mais curtos entre os nós de uma rede que passam por essa ligação (Oliveira & Gama, 2010).

Proximidade é uma medida que permite descobrir a posição de um nó numa rede, isto é, o quão perto está este nó dos restantes nós da rede. Esta medida é dada pela média de todos os caminhos mais curtos entre um nó e os restantes nós da rede e é representada pela fórmula:

$$Cl_v = \frac{n-1}{\sum_{u \in V(G)\setminus v} d(u,v)} \qquad (2.5)$$

No contexto de redes sociais, **proximidade** permite medir quão rápido consegue um nó chegar a qualquer outro nó da rede (Oliveira & Gama, 2010).

Centralidade *eigenvector* mede a importância de um nó numa rede tendo em conta as suas ligações com outros nós, atribuindo-lhes uma pontuação relativa de acordo com os nós a que estes estão ligados. Em resumo, permite verificar se o nó tem uma ligação forte a outros nós bem ligados da rede (Oliveira & Gama, 2010).

Esta medida que é obtida através do primeiro ***eigenvector*** da matriz de adjacência da rede é dada pela seguinte fórmula:

$$x_i \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j \qquad (2.6)$$

Coeficiente de aglomeração local é uma medida que permite identificar a transitividade da vizinhança de um dado nó. Por transitividade entende-se o nível de coesão entre os vizinhos de um dado nó. Esta medida é bastante usada na análise de redes sociais uma vez que estas são naturalmente transitivas, ou seja, a probabilidade dos amigos de um actor serem amigos entre si é elevada. Este coeficiente é dado pela seguinte fórmula:

$$c_i = \frac{2|e_{jk}|}{k_i(k_i-1)} : \; v_j, v_k \in N_i, e_{jk} \in E \qquad (2.7)$$

Onde N_i é a vizinhança do nó v_i , e_{jk} é o vértice que liga v_j a v_k , k_i é o **grau** do nó v_i e $|e_{jk}|$ é a proporção de ligações entre os nós da vizinhança do nó v_i (Oliveira & Gama, 2010).

2.3.2 Medidas ao nível da rede

As medidas ao nível da rede permitem identificar características da estrutura global da rede e do comportamento que a gerou.

O **diâmetro** da rede é dado pelo valor máximo de excentricidade dos nós da rede enquanto que o **raio** da rede é obtido através do valor mínimo. Estas medidas são dadas pelas seguintes fórmulas (Oliveira & Gama, 2010).

$$D = \max\{\in_v : v \in V\} \quad (2.8)$$

$$R = \min\{\in_v : v \in V\} \quad (2.9)$$

A **distância geodésica média** permite obter uma indicação do grau de afastamento dos nós (em média). A fórmula que permite calcular esta métrica está representada em baixo.

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d(i,j) \quad (2.10)$$

No caso da rede ser não conectada terá que ser utilizada outra fórmula uma vez que o facto de não existir ligações entre grupos de nós eleva o valor da fórmula a ∞. Sendo assim a fórmula para o cálculo da **distância geodésica harmónica média** é dada por (Oliveira & Gama, 2010):

$$l^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} \frac{1}{d(i,j)} \quad (2.11)$$

Grau médio é a média do **grau** de todos os nós da rede pelo que pode ser usada para medir a conectividade global da rede (Oliveira & Gama, 2010).

$$k = \frac{1}{n} \sum_{i=1}^{n} k_i \quad (2.12)$$

Reciprocidade é uma medida que é usada especificamente para redes direccionadas e que permite medir a tendência de pares de nós formarem ligações mútuas entre si. Apesar de existirem várias formas de calcular esta medida a mais utilizada é calculando o rácio entre número de ligações mútuas na rede e o número total de ligações.

$$r = \frac{\#mut}{\#mut+\#assim}, \quad 0 < r < 1 \qquad (2.13)$$

onde *#mut* representa o número de pares mútuos (ligação ente dois nós nos dois sentidos) e *#assim* o número de pares assimétricos (ligação entre dois nós apenas num sentido) (Oliveira & Gama, 2010).

Densidade é uma medida que permite explicar o nível geral de ligações numa rede. Esta métrica é obtida pelo quociente do número de ligações da rede sobre o número máximo possível de ligações. Para redes sem ligações assume o valor 0 enquanto que para redes completamente ligadas assume o valor de 1. Redes completamente conectadas também podem ser designadas de grafos completos ou cliques. A fórmula de cálculo da **densidade** é a seguinte:

$$\rho(G) = \frac{m}{m_{max}} \qquad (2.14)$$

Coeficiente de aglomeração médio, em analogia com o seu homólogo local, permite calcular a transitividade de toda a rede. Apesar de existirem várias formas de fazer este cálculo a adoptada neste trabalho foi proposta por Watts e Strogatz em que o **coeficiente de aglomeração médio** é dado pela média dos **coeficientes de aglomeração locais** de todos os nós da rede. A fórmula de cálculo é a seguinte:

$$c = \frac{1}{n}\sum_i c_i \qquad (2.15)$$

(Oliveira & Gama, 2010) (Watts & Strogatz, 1998).

Um valor elevado do coeficiente de aglomeração médio é frequentemente observado em redes "pequeno mundo", significando que existe uma probabilidade elevada de formação de cliques (Oliveira & Gama, 2010).

2.4 Modelos de Redes

Na secção anterior foram explicadas medidas estatísticas utilizadas no estudo de redes. No entanto, a análise de redes também pode ser feita através da sua topologia. O modelo mais simples e conhecido de redes é o **grafo aleatório**. No entanto, dada a sua simplicidade, este modelo não é por si só suficiente para compreender o comportamento de uma rede. Outro modelo bastante utilizado é o modelo "pequeno mundo".

2.4.1 Grafo aleatório

O modelo **grafo aleatório** é caracterizado pela colocação aleatória de ligações com uma probabilidade associada entre um número fixo de nós. Nesta rede, as $\frac{1}{2}n(n-1)$ ligações possíveis têm uma probabilidade p associada e o número de ligações em cada nó é distribuído de acordo com uma distribuição binomial ou de Poisson (Newman, 2003). Quando $p = 0$, é gerado um grafo de ordem perfeita enquanto que para $p = 1$ se obtém um grafo completamente caótico. Na Figura 2.4 podem-se observar dois grafos aleatórios com diferentes parâmetros de p (Gama & Oliveira, 2010) (Newman, 2003).

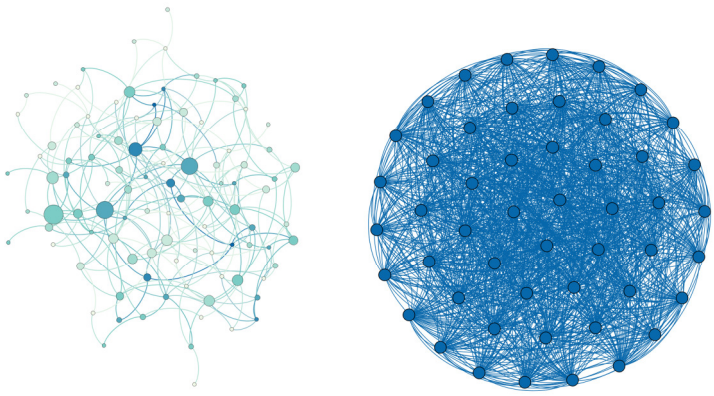

Figura 2.4: Exemplo de dois *random graphs* p=0.04 e p=1

2.4.2 Modelo “pequeno mundo”

Uma rede “pequeno mundo” é uma rede caracterizada por um número elevado de agrupamentos e uma grande proximidade entre os nós. Este modelo assenta no pressuposto em que a maioria dos nós não são vizinhos entre si mas conseguem comunicar através de um número pequeno de ligações (Newman, 2003).

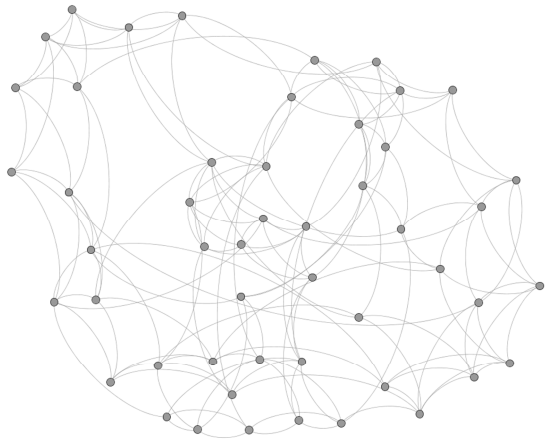

Figura 2.5: Rede “pequeno mundo” (modelo Watz-Strogatz)

2.5 Propriedades das Redes

As redes do mundo real são redes não aleatórias. Este facto deve-se à simples razão de existir sempre um motivo para a criação ou formação da rede. Nos seguintes parágrafos são descritas propriedades que são normalmente observadas em redes com diferentes tipos de estrutura e que podem servir como base para a explicação de comportamentos de redes no mundo real.

Efeito "pequeno mundo"

O efeito de **"pequeno mundo"** foi observado pela primeira vez por Stanley Milgram na sequência de várias experiências que realizou. Milgram enviou cartas a várias pessoas pedindo que estas reenviassem a carta até esta chegar a um determinado destinatário. Apesar de na prática não ter construído nenhuma rede, Milgram conseguiu demonstrar que é possível atingir um determinado indivíduo em poucos passos. No caso em questão foi demonstrado que o comprimento médio dos percursos que efectivamente atingiram o destino foi de aproximadamente seis.

Na realidade esta propriedade é bastante importante uma vez que demonstra que é possível contactar, através de uma rede de conhecimentos, praticamente qualquer pessoa no planeta.

Em termos matemáticos o efeito de **"pequeno mundo"** indica que a **distância geodésica média** entre pares de nós tem uma relação logarítmica (ou menor) com o tamanho da rede, ou seja, o número de nós da rede influencia muito pouco a distância geodésica média**.** (Gama & Oliveira, 2010) (Newman, 2003).

Transitividade ou clusterização

De acordo com Newman, **transitividade** ou **clusterização** são propriedades de redes em que existem ligações triangulares. Verifica-se que, se numa rede, um nó A estiver ligado a um nó B e o nó B estiver ligado a um nó C, então a probabilidade de A estar ligada a C é alta. O equivalente desta propriedade para as redes sociais é: o amigo de um amigo, provavelmente também é meu amigo. Newman propõe duas formas de calcular esta métrica. Na primeira (eq. 2.15) C mede a fracção de triplos que tem a terceira ligação que completa o triângulo. Assim sendo, C é a probabilidade média de dois nós de uma rede que são vizinhos do mesmo nó serem vizinhos.

$$C = \frac{3 \times \text{número de triangulos na rede}}{\text{número de triplos de nós ligados}} \quad (2.16)$$

Na segunda (eq 2.16 e 2.17) é calculado o coeficiente por nó e posteriormente é feita a média (Watts & Strogatz, 1998).

$$c_i = \frac{\text{número de triangulos ligados ao nó } i}{\text{número de triplos centrados no nó } i} \quad (2.17)$$

$$C = \frac{1}{n}\sum_i c_i \quad (2.18)$$

Estas duas abordagens não são equivalentes porque enquanto que uma calcula o rácio da média a outra calcula a média do rácio. Neste trabalho será utilizada a segunda uma vez que é a utilizada em cálculo computacional (confirmar se é mesmo este modelo usado no Gephi) (Newman, 2003) (Watts & Strogatz, 1998).

Distribuição de probabilidade do grau

A **distribuição de probabilidade** p_k é a distribuição de probabilidades do grau dos nós de toda a rede. p_k pode também ser definido pela fracção de nós na rede que tem grau *k*. Desta forma, se a rede tiver *n* nós e n_k destes nós tiverem grau *k*, então para este valor do grau tem-se uma probabilidade $P(k) = \frac{n_k}{n}$. Fazendo este cálculo para cada *k* de um total de *K* graus na rede, obtém-se a distribuição de probabilidade do grau da rede (Gama & Oliveira, 2010) (Newman, 2003).

Resiliência da rede

Esta propriedade das redes consiste na resiliência de uma rede relativamente à remoção dos seus vértices. Permite reflectir o impacto na conectividade da rede e serve como um indicador da coesão da rede. Quando são removidas ligações numa rede o comprimento

médio dos caminhos que ligam a rede aumenta e dependendo da forma como é feita esta remoção o efeito na rede pode ser bastante diferente. Enquanto que a maioria das redes apresentam alguma robustez à remoção aleatória de ligações, quando é feita uma remoção localizada de ligações com um grau alto o impacto na rede é bastante significativo (Gama & Oliveira, 2010) (Newman, 2003).

Homofilia

A análise de redes e da sua estrutura, muitas vezes passa por tentar identificar que nós se agrupam entre si. Na maioria das redes existem sempre alguns tipos diferentes de nós em que, dada a sua natureza, podem ter uma probabilidade mais elevada de se ligar a outros tipos de nós ou a nós do mesmo tipo. Como exemplo pode-se considerar uma rede de um ecossistema em que os nós representam plantas, animais herbívoros e animais carnívoros e as ligações identificam uma relação de subsistência. Nesta rede existem bastantes ligações entre plantas e herbívoros e herbívoros e carnívoros mas certamente muito poucas a ligar herbívoros a herbívoros e carnívoros a plantas. Um exemplo para redes tecnológicas, mais especificamente a internet, dado por Maslov et al. identifica três grandes tipos de nós: fornecedores ISP de alto nível/*backbone* globais, fornecedores ISP locais e utilizadores finais. Aqui existirão bastantes ligações entre os ISPs e os utilizadores finais, entre os ISP e os fornecedores do *backbone* e muito poucas entre ISPs e ISPs e fornecedores de *backbone* e utilizadores finais (Newman, 2003) (Maslov & Zaliznyak, 2002).

Em redes sociais, este tipo de ligação selectiva é denominada de **homofilia** ou **mistura**. Um exemplo clássico é a mistura de casais entre diferentes raças (brancos, hispânicos, negros) (Gama & Oliveira, 2010) (Newman, 2003).

Correlação entre graus

Correlação entre graus é usado para classificar casos específicos de **homofilia** em que a ligação é feita de acordo com uma propriedade escalar do vértice. Esta propriedade surge na necessidade de se saber se, numa dada rede, os vértices com um grau alto preferem associar-se a outros vértices com grau alto ou a vértices com um grau baixo. Na prática tem-se observado os dois tipos de situação. Esta propriedade permite tirar conclusões bastante interessantes (Newman, 2003).

Estrutura em comunidade

Na maioria das redes sociais reais, é possível observar estruturas semelhantes a comunidades. Por outras palavras, é possível encontrar conjuntos de nós bastante interligados entre si e com poucas ligações a outros conjuntos de nós também densamente ligados. Esta característica é bastante fácil de explicar uma vez que faz parte da experiência da vida real em que as pessoas têm a tendência de se agrupar de acordo com algum tipo de interesse, faixa etária, ocupação, etc. Na secção 2.6 será feita uma descrição detalhada de alguns métodos utilizados na detecção de comunidades.

Navegação na rede

A experiência efectuada por Stanley Milgram que consistiu em enviar cartas de pessoa para pessoa permitiu identificar o efeito **"pequeno mundo"**. No entanto, esta experiência permitiu também retirar outra conclusão, mais tarde identificada por Kleinberg, que para além do facto de existirem caminhos curtos através de uma rede entre indivíduos aparentemente distantes, estes caminhos curtos são facilmente descobertos por indivíduos comuns. Se esta propriedade identificada para redes sociais fosse possível aplicar a redes tecnológicas (artificias), então estas redes poderiam ser

usadas para construir estruturas bastante mais eficientes como bases de dados ou redes de computadores *peer-to-peer* (Newman, 2003) (Kleinberg, 2000).

Outras propriedades e características de redes

Para além das propriedades de redes identificadas anteriormente existem outras que também tem sido alvo de alguma atenção.

O tamanho do maior componente, por exemplo, numa rede de comunicação como a Internet, representa a maior fracção da rede em que a comunicação é possível, sendo assim uma medida de eficiência da rede em "fazer o seu trabalho". Este componente é muitas vezes equiparado ao conceito teórico de "componente gigante". O tamanho do segundo maior componente é também é medido em algumas situações (Newman, 2003).

No decorrer deste trabalho também serão referidas algumas características das redes ou grafos, nomeadamente para grafos cíclicos/acíclicos, conectados e cliques.

Grafos **cíclicos** são grafos que contem circuitos fechados de nós ou anéis. Pelo contrário, grafos **acíclicos** não são caracterizados por este tipo de ligações (ex: árvores) (Freeman, 1979) (Gama & Oliveira, 2010).

Um grafo é **conectado** se existe um caminho entre qualquer par de nós de uma rede e **desconectado** se esta condição não se verificar (Gama & Oliveira, 2010).

Clique ou grafo completo é uma designação atribuída a uma rede completamente conectada. Esta rede tem um valor de densidade igual a 1.

2.6 Comunidades

O conceito de comunidade é usado no contexto de redes de uma forma semelhante do que no mundo real, ou seja, dependendo de factores comuns podem-se identificar

grupos de nós bastante próximos. Estes factores podem ser tão simples como faixa etária, pertença a um grupo, localização geográfica, etc. Numa rede, as comunidades podem ser identificadas por um conjunto de ligações homogéneas entre um grupo de nós. Adicionalmente, uma comunidade também é caracterizada por um baixo número de ligações para fora deste grupo de nós. São, pois caracterizadas por uma densidade alta de ligações no interior da comunidade e baixa no exterior.

2.6.1 Detecção de comunidades

Uma comunidade é caracterizada por um grupo de nós densamente conectados que se distinguem dos restantes. Este tipo de estrutura é bastante observado e tem sido objecto de estudo devido à sua importância. A descoberta de comunidades, das razões subjacentes à sua constituição e a caracterização dos elementos nelas presentes permitem tirar conclusões sobre uma rede que não seriam possíveis através de outros tipos de análise.

Existem vários algoritmos aplicados à detecção de comunidades como por exemplo aglomeração hierárquica, modelação de blocos, Girvan-Newman, Blondell, entre outros. Nos seguintes parágrafos é feita uma descrição dos métodos de Girvan-Newman e Blondell.

2.6.1.1 Método de Girvan-Newman

O método de Girvan-Newman consiste na eliminação iterativa de ligações tendo por base medidas de intermediação. Recorde-se que a intermediação permite medir a quantidade de informação que passa por um nó ou ligação entre diferentes grupos. Após cada iteração estas medidas são novamente calculadas para o passo seguinte. O modo de acção deste método é o de fazer uma decomposição faseada da rede até obter um dado

número de comunidades. O critério de paragem é obtido através do cálculo de uma métrica indicativa da força estrutural de uma comunidade.

O algoritmo proposto por estes autores funciona da seguinte forma:

1 – Cálculo das medidas de intermediação para todas as ligações da rede;

2 – Procurar e remover a ligação com o maior valor de intermediação;

3 – Recalcular as medidas de intermediação para as restantes ligações da rede;

4 – Repetir a partir do 2º passo.

Como medidas de cálculo de intermediação, estes autores propõem: "medição dos caminhos geodésicos mais curtos", "intermediação de caminho aleatório (*random-walk*)" e, derivada da teoria dos circuitos, "intermediação de fluxo de corrente".

Para a quantificação da força dentro das comunidades resultantes de cada iteração, sendo que este será o critério de paragem do algoritmo, é proposto o conceito de **modularidade**. Esta medida, apresentada em baixo, é baseada na medida de **homofilia** proposta por Newman.

$$Q = \sum_i (e_{ii} - a_i^2) \qquad (2.19)$$

em que $Tr\ e = \sum_i e_{ii}$ é a fracção de ligações na rede que ligam nós na mesma comunidade, sendo que esta medida deve ser alta se houver uma boa divisão de comunidades. $a_i = \sum_j e_{ij}$ representa a fracção de ligações que se ligam a vértices na comunidade i (Newman, 2003) (Girvan & Newman, 2004).

2.6.1.2 Algoritmo de Blondel

Este método propõe uma abordagem heurística baseada na optimização da modularidade da rede.

Uma característica interessante deste método é o facto de não ter como limitação a capacidade de processamento computacional (evidente nos restantes métodos de detecção de comunidades) mas sim a capacidade de armazenamento.

O algoritmo é dividido em duas fases sendo que na primeira é feito um agrupamento de nós em comunidades e na segunda, as comunidades são transformadas em nós. A rede obtida na segunda fase é novamente sujeita aos cálculos efectuados na primeira fase. Este processo termina quando já não se conseguem detectar mais alterações e se obteve um valor máximo para a modularidade.

Primeira fase:

1 – Assignar uma comunidade para cada nó da rede;

2 – Para cada nó i, tendo em conta os seus j vizinhos, é calculado o ganho de modularidade caso i tenha sido eliminado da sua comunidade e se passasse a fazer parte da comunidade de j;

3 – i é colocado na comunidade para a qual o ganho é máximo, Se o ganho não for positivo, i mantém-se na sua comunidade inicial;

4 – Repetir passos 2 e 3 até não haver melhoria de modularidade.

Segunda Fase:

1 – Reconstrução da rede em que os nós são as comunidades encontrada na primeira fase. Neste caso o peso das ligações entre os novos nós são dados pela soma dos pesos das ligações dos nós das duas comunidades;

2 – Repetição da primeira fase.

O cálculo da modularidade é feito recorrendo à fórmula (2.20).

$$\Delta Q = \left[\frac{\sum in + k_{i,in}}{2m} - \left(\frac{\sum tot + k_i}{2m}\right)^2\right] - \left[\frac{\sum in}{2m} - \left(\frac{\sum tot}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \qquad (2.20)$$

em que $\sum in$ representa a soma dos pesos das ligações dentro de uma comunidade C, $\sum tot$ é a soma dos pesos das ligações incidentes nos nós de C, k_i é a soma dos pesos das ligações incidentes no nó i, $k_{i,in}$ é a soma dos pesos das ligações de i aos nós de C e m é a soma dos pesos de todas as ligações na rede (Blondel et al., 2008).

Pode-se observar a aplicação do método de Blondel na Figura 2.6. Na imagem da esquerda pode-se observar a rede inicial. Na imagem central é visível a classificação por cor de cada nó, de acordo com a sua modularidade. A imagem da direita representa o agrupamento de cada comunidade em novos nós.

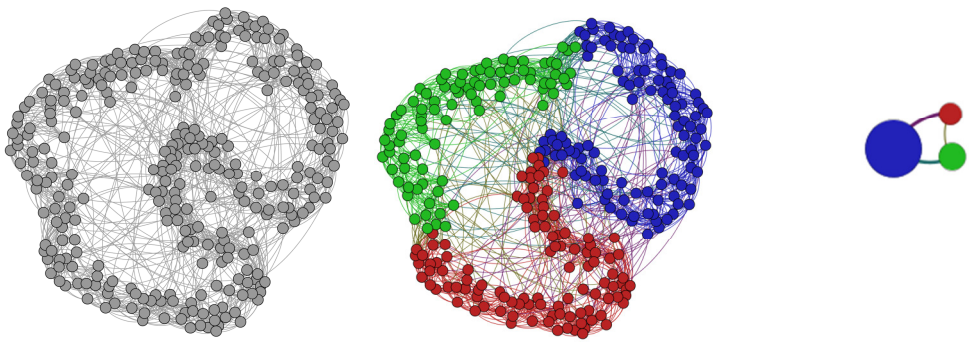

Figura 2.6: Aplicação algoritmo Blondel. Esquerda: Rede inicial; Centro: Cálculo de modularidade; Direita: Agrupamento comunidades

2.7 Análise de redes dinâmicas

2.7.1 Introdução às redes dinâmicas

Na literatura podem-se encontrar algumas definições distintas para a análise de redes dinâmicas. Enquanto alguns autores abordam o tema de redes dinâmicas apenas pela introdução da componente temporal (Yu-Ru *et al.* 2007) (Asur et al. 2007), outros consideram DNA (*Dynamic Nerwork Analysis*) bastante mais abrangente (Carley, 2003). Na abordagem defendida por Carley, a rede é considerada multi-modo, multi-ligação e multi-nível. Multi-modo significa que existem vários tipos de nós, como por exemplo, pessoas e locais. Multi-ligação é referente à possibilidade de existirem vários tipos de ligação na mesma rede, como amizade, ligação comercial ou compra de um mesmo produto. Multi-modo significa que é possível haver nós membros de outros nós, ou seja, um colaborador que pertence a uma empresa (Carley, 2003).

Uma base comum em todas estas definições é a variante temporal. Neste trabalho, será apenas esta a variável considerada na análise dinâmica. Nesta óptica, a questão dos instantes temporais é abordada na forma de janela temporal em dois formatos: janela deslizante e janela acumulada.

2.7.2 Detecção de comunidades em redes dinâmicas

Na detecção de comunidades também podem ser encontradas formas distintas de análise da evolução temporal.

Alguns autores propõem uma análise da evolução de comunidades através de fotografias temporais e fazendo a comparação evolutiva em instantes consecutivos (Asur et al. 2007). Chi et al., numa versão mais avançada, apresenta um método de análise temporal de evolução de agrupamentos com uma componente de amortecimento temporal, isto é, a escolha das divisões dos vários agrupamentos é decidida de forma a

afectar o mínimo possível a consistência temporal. Uma abordagem diferente, proposta por Yu-Ru *et al.* considera a introdução do conceito de comunidade *soft* em que se permite que um dado nó pertença a duas comunidades distintas no mesmo instante temporal.

Uma vez que o objectivo deste trabalho não é apenas fazer a análise de evolução de comunidades, a detecção de comunidades é efectuada de uma forma bastante simplificada recorrendo ao proposto por Asur et al.

Desta forma, e de acordo com este autor, a detecção da alteração do estado das comunidades em intervalos temporais consecutivos é efectuada recorrendo à análise de eventos.

Os eventos propostos são **continuação**, **k-fusão**, **k-separação**, **formação** e **dissolução**.

Um agrupamento C_{i+1}^{j} é considerado como a **continuação** de C_i^k se V_{i+1}^{j} é igual a V_i^k.

K-fusão ocorre entre dois agrupamentos diferentes C_i^k e C_i^l se no período seguinte existir um agrupamento que contenha $k\%$ dos nós pertencentes a estes dois agrupamentos.

Por outro lado, **k-separação** ocorre para um agrupamento C_i^l s $k\%$ dos nós deste agrupamento estão presentes em dois agrupamentos diferentes no instante temporal seguinte.

Formação é atribuído a um agrupamento novo, C_{i+1}^k, se não existir nenhum conjunto de pares de nós que tenham pertencido a algum agrupamento no intervalo anterior.

Um agrupamento C_i^k, considera-se como **dissolvido** se nenhum conjunto de pares de nós deste agrupamento exista num agrupamento no intervalo seguinte.

Capítulo 3

3 Metodologia

Neste capítulo é feita a apresentação da metodologia proposta para a análise dos dados provenientes de uma base de dados transaccional de uma aplicação CRM *(Customer Relationship Management)* à luz dos conceitos de análise de redes sociais. Os dados são importados para uma aplicação que permita tanto a visualização gráfica da rede como a realização de várias medições ao nível dos nós, da rede, detecção de comunidades e evolução temporal.

3.1 Representação da Rede

A representação da rede é feita sendo os nós os clientes e as ligações os produtos comprados. Sendo assim, um cliente fica ligado a outro se ambos compraram o mesmo produto. Nesta construção, é ignorada a direcção da ligação, ou seja, se um cliente comprou o produto primeiro que o cliente seguinte.

3.2 Análise da Rede

A análise da rede é constituída por várias fases. Na primeira fase são analisadas e interpretadas várias métricas associadas aos nós e à rede. Na fase seguinte, recorrendo ao algoritmo de Blondell, é realizada a detecção de comunidades. As duas fases referidas anteriormente são aplicadas recursivamente a duas dinâmicas temporais que, segundo esta metodologia, consistem numa janela deslizante e numa janela cumulativa. Os últimos passos da metodologia compreendem a análise de cada uma das dinâmicas temporais e uma comparação final dos dois métodos.

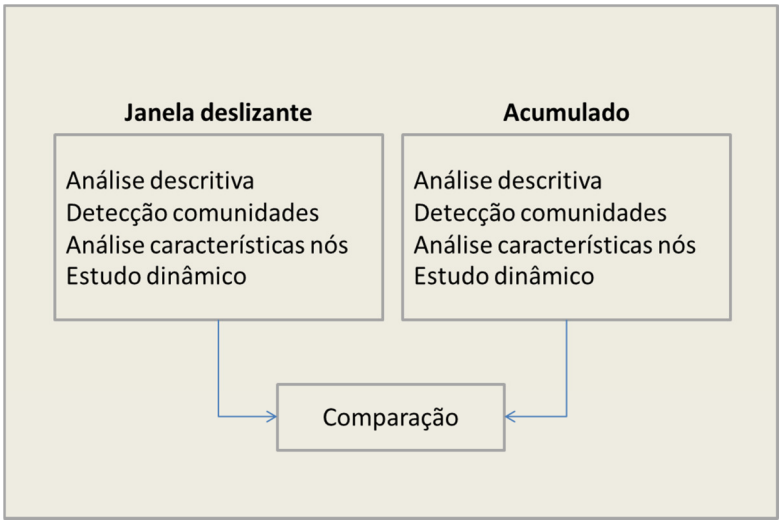

Figura 3.1: Metodologia proposta

A interpretação dos vários resultados de acordo com cada tipo de análise e tendo como óptica o objecto de estudo é descrita nos seguintes parágrafos.

3.2.1 Métricas ao nível dos nós

Ao nível dos nós são analisadas a métricas de **grau**, **intermediação e centralidade *eigenvector*.**

A medida do **grau** permite analisar se existem muitas ligações para um determinado nó da rede. Aplicando este conceito ao tema deste trabalho, pode-se concluir que um cliente que tenha um grau bastante elevado é um cliente que compra diferentes tipos de produtos à empresa. Por outro lado, um cliente com um grau baixo é caracterizado por adquirir uma baixa variedade de produtos. Esta métrica, por si só, não permite auferir directamente a relevância do cliente para a empresa, se bem que o contrário possa ser demonstrado para uma rede social. Analisando o **grau médio** pode-se tirar uma conclusão relativa à estrutura de clientes, ou seja, a partir de um valor alto do grau médio conclui-se que a rede é bastante densa, caracterizada por um conjunto elevado de

clientes pluri-produto. Um valor baixo, revela naturalmente o contrário. Sendo uma média, valores para além dos próximos dos extremos não permitem concluir nada em concreto uma vez que podem existir zonas da rede bastante conectadas em detrimento de outras pouco conectadas.

A medida da **intermediação** é efectuada ao nível dos nós e permite identificar se um cliente ocupa uma posição crítica na rede, servindo de ponte entre grupos distintos de outros clientes. Em redes sociais, estes agentes são denominados de *gatekeepers*. Um valor elevado desta medida pode levar a várias conclusões:

- Cliente está a transitar de negócio;

- Cliente está a mudar de tecnologias;

- Cliente está a diversificar;

- O próprio modelo do negócio do cliente abrange várias áreas/produtos.

Relativamente às várias alternativas apresentadas, não se considera possível escolher uma sem a interpretação da evolução temporal ou análise do modelo de negócio da empresa cliente.

A **média da intermediação** permite auferir, de uma forma global, se existem muitos clientes com posições de *gatekeeper* ou de transição.

Em conjunto com a análise temporal a medida da intermediação é bastante interessante, uma vez que pode permitir identificar tendências na rede de clientes que podem, por sua vez, revelar tendências do mercado em geral.

Finalmente, a **centralidade *eigenvector*** permite inferir a importância de um cliente na rede tendo como base as ligações dos nós a que ele próprio está ligado. Em redes sociais esta medida identifica os nós com a melhor rede de contactos. Aplicando este conceito a uma rede de clientes, esta medida permite identificar clientes que compram os mesmos produtos que outros clientes que, por sua vez, também compram uma grande variedade de produtos. Estes clientes poderão ser considerados clientes ”montra” dada a sua

visibilidade na rede e, regra geral, podem ser os primeiros a ser abordados em estratégias de marketing para introdução de novos produtos.

A **centralidade *eigenvector* média** não é muito útil uma vez que, sendo uma média os resultados e suas interpretações podem ser bastante diferentes consoante o tipo de rede. Considera-se mais sensato analisar esta medida individualmente para os clientes.

3.2.2 Métricas ao nível da rede

Ao nível da rede são analisadas as medidas de diâmetro, raio, densidade, número de percursos mais curtos, comprimento médio de percursos, coeficiente de aglomeração médio e modularidade.

O diâmetro permite dar uma indicação sobre a proximidade entre pares de clientes na rede. Um diâmetro elevado é sinónimo da existência de uma diversidade elevada de produtos que não é partilhada por alguns clientes, ou seja, os clientes que protagonizam a medida do diâmetro (pelo menos estes), compram produtos diferentes entre si.

A medida da densidade da rede, variando entre 0 e 1 permite determinar o nível geral de ligações na rede. Um valor alto demonstra que existe uma grande proximidade entre clientes significando que existe uma grande homogeneidade nos produtos adquiridos. Por outro lado, um valor baixo da densidade revela que as escolhas dos clientes são bastante diferentes. Esta métrica pode ser interpretada da seguinte forma: uma rede de clientes densa demonstra que existe uma certa maturação, tanto a nível dos clientes como ao seu comportamento na compra de produtos; por outro lado, uma rede pouco densa revela uma rede menos madura, uma vez que os clientes têm preferências diferentes no que concerne o tipo de produto. Desta interpretação pode-se inferir que poderá fazer sentido tentar abordar novos clientes ou vender novos produtos para redes densas, ou partir para uma estratégia de convergência para redes pouco densas.

O comprimento médio de percursos indica a distância média na rede entre todos os pares de clientes. Desta forma, um valor alto pode indicar que existem poucas ligações

ou que existem bastantes *hubs*. Um valor baixo pode significar que os clientes estão bastante conectados, de uma forma geral.

A medição do coeficiente de aglomeração serve como um indicador para ligações triádicas ou formação de cliques. O efeito prático de um valor alto desta métrica é a existência de conjuntos de clientes que compram os mesmos produtos ou, em que pelo menos dois a dois compram os mesmos produtos, formado um circuito triangular entre eles.

A métrica da modularidade tem um significado bastante directo uma vez que permite descobrir se a rede de clientes apresenta, de facto, a existência de comunidades.

3.2.3 Detecção de comunidades

A detecção de comunidades é efectuada com a aplicação do algoritmo de *Blondell* baseado na optimização da métrica de modularidade.

O resultado da aplicação deste método consiste no agrupamento dos vários clientes em comunidades com a identificação da taxa de representatividade para rede global de cada uma. As comunidades formadas são assim caracterizadas por um grande número de ligações entre os seus elementos significando que os clientes pertencentes a uma comunidade compram o mesmo ou os mesmos produtos.

3.2.4 Dinâmica temporal

A aplicação da dinâmica temporal permite extrair um conhecimento adicional da rede. Com a evolução temporal podem-se identificar alterações no comportamento de clientes nomeadamente transições de negócio, mudanças de tecnologia ou mesmo diversificação. Para além destes comportamentos também poderá ser possível identificar tendências de mercado e de maturação de clientes.

Nesta metodologia são propostas duas abordagens para a análise temporal:

- Análise com janela deslizante;

- Análise com janela cumulativa;

Na Figura 3.2 pode-se observar do lado esquerdo um exemplo de janela deslizante com um tamanho de janela de três períodos e uma sobreposição de dois períodos entre intervalos. Do lado direito, está representada uma dinâmica temporal com janela acumulada em que a janela temporal é cada vez maior ao longo do tempo.

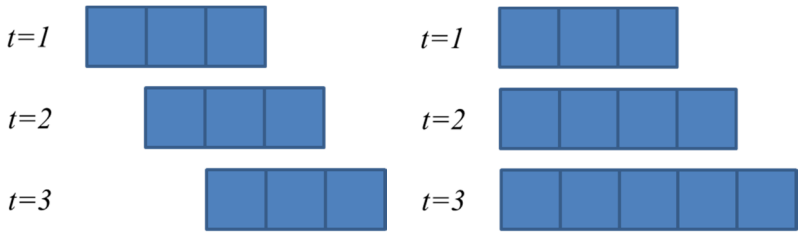

Figura 3.2: Janela deslizante (esquerda); janela acumulada (direita)

Dependendo do método de análise, os resultados obtidos são diferentes e complementares.

O método de janela deslizante permite introduzir a variável de envelhecimento, em que clientes antigos deixam de ser considerados nas medições e análise em cada instante. Como configuração da janela, propõem-se que a janela tenha o tamanho de três períodos e que o deslocamento entre janelas seja de uma unidade temporal. Desta forma garante-se que existe sempre um período em comum entre janelas consecutivas permitindo uma transição menos disruptiva.

Por outro lado, a janela cumulativa, ao manter o histórico de ligações antigas tem sempre em consideração a influência e características de todos os clientes.

Dependendo do modelo de negócio de uma empresa um método poderá ser mais vantajoso do que o outro uma vez que para empresas que actuam em contextos bastante voláteis e com uma grande dinâmica, o comportamento da rede de clientes é completamente diferente de empresas em que actuam em contextos menos dinâmicos.

Neste método são aplicadas as duas dinâmicas sendo que posteriormente serão comparados os resultados obtidos em cada uma.

3.2.5 Dinâmica temporal e detecção de comunidades

Como demonstrado no Capítulo 2, a dinâmica temporal traz algumas questões relativamente à análise e detecção de comunidades.

Estas questões prendem-se com a identificação das comunidades, com o descobrir se uma comunidade apresentada num dado intervalo é a mesma no intervalo seguinte, se desapareceu, se fundiu a outra ou se deu origem a outras comunidades.

Nesta metodologia é proposta a classificação de acordo com o modelo de Asur et Al. com ligeiras modificações. Neste caso, considera-se que é permitido a uma comunidade continuar, mesmo que não mantenha todos os seus elementos. Sendo assim, por um lado possível a uma comunidade, continuar e separar-se ao mesmo tempo enquanto por outro é possível desaparecer e fazer uma separação ao mesmo tempo.

Em cada instante temporal são analisados os vários agrupamentos e é feita a identificação do evento que influenciou cada um destes agrupamentos para o instante temporal seguinte. Desta forma, consegue-se fazer o desenho evolutivo das comunidades como se pode ver na Figura 3.3.

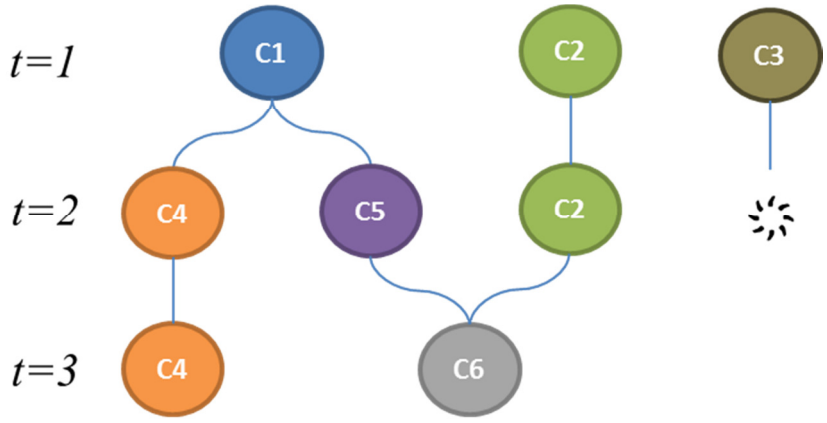

Figura 3.3: Evolução comunidades

Este desenho permite ter uma visão bastante interessante da evolução das comunidades. O mapeamento das comunidades e da sua interacção ao longo do tempo permite identificar tendências e fazer surgir algumas perguntas que façam sentido do ponto de vista comercial.

Tendo como exemplo a Figura 3.3, podem ser feitas as seguintes questões:

1 – O que levou a comunidade C3 a desaparecer?

2 – A separação da comunidade C1 em duas, C4 e C5, a que se deve?

3 – Qual a razão para a fusão da comunidade C5 com a C2?

Outra questão que pode surgir e que não está representada na figura deriva do facto de existir uma comunidade que se mantém ao longo do tempo.

As respostas para estas questões podem ser diversas, e dependerão certamente de negócio para negócio mas identificam sempre um tipo específico de comportamento por parte dos clientes.

No caso apresentado, faria sentido tentar perceber qual a razão que levou a desaparecer a comunidade C3. Pode ter sido devido a algum produto ter sido descontinuado. Sendo assim, oferecer um produto equivalente aos mesmos clientes ou então tentar vender-lhes outro produto poderia evitar que deixassem de fazer parte da rede (pelo menos nesse instante). Esta interpretação poderia ser usada para responder ao facto de C2 se fundir a

C5, ou seja, C2 pode não ter desaparecido porque se conseguiu que consumissem os mesmos produtos que a comunidade C5, dando origem a C6.

Em resumo, as interpretações para o comportamento podem ser diversas e as respostas só serão possíveis analisando detalhadamente as características das comunidades à vista da estratégia da empresa.

Capítulo 4

4 Estudo de Caso

Neste Capítulo é aplicada a metodologia descrita no Capítulo três a um caso real de uma rede clientes proveniente de uma aplicação CRM em que o modelo de negócio é B2B (*business to business).*

4.1 Análise da rede

A análise da rede pode ser efectuada com o recurso a qualquer aplicação que permita fazer as medições propostas. Neste caso foi utilizada a aplicação *Gephi*.

O estudo efectuado está separado em duas seções, uma para análise temporal com janela deslizante e outra com janela temporal acumulada num período total de doze meses. Em cada uma destas secções são analisadas as métricas em cada instante temporal.

4.2 Análise janela deslizante

4.2.1 Apresentação visual

Com a aplicação do método de janela deslizante obteve-se a seguinte representação da rede nos vários instantes.

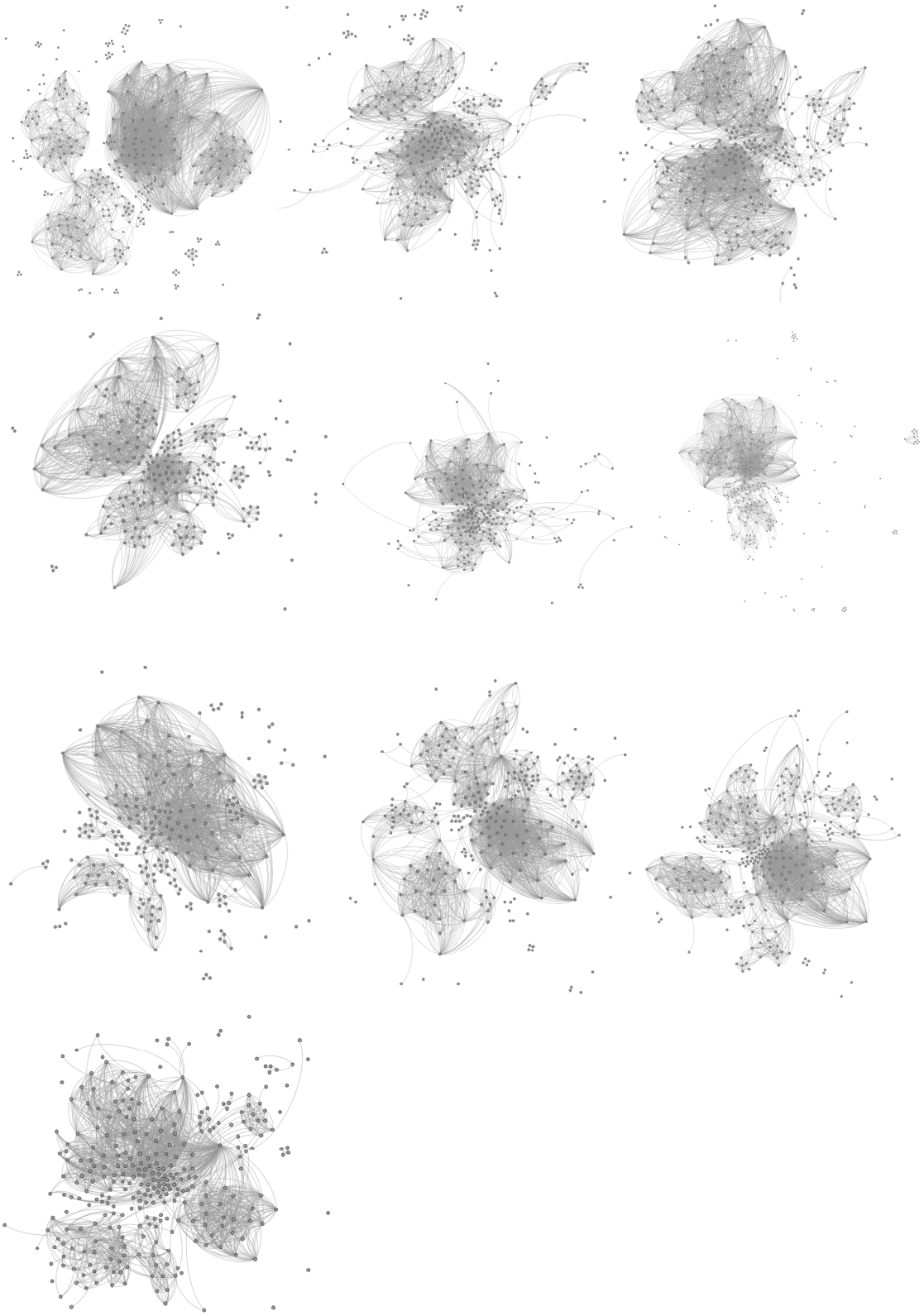

Figura 4.1: Janela deslizante

Pela análise visual das imagens da rede de clientes facilmente se observa a descontinuidade provocada pela janela deslizante. Apesar deste facto, existem algumas zonas da rede que mantém alguma consistência ao longo do tempo. O facto da descontinuidade ser bastante visível pode significar que a rede em análise não é suficientemente dinâmica para os parâmetros escolhidos para a janela deslizante.

Relativamente ao número de nós e ligações em cada período da janela, estes estão representados na Tabela 4.1.

Intervalo	1-3	2-4	3-5	4-6	5-7	6-8	7-9	8-10	9-11	10-12
Nós	336	302	296	260	227	300	246	288	303	275
Ligações	3509	1957	2611	1951	1519	2398	1858	2383	2359	1775

Tabela 4.1: Número de nós e ligações

4.2.2 Medidas ao nível dos nós

Ao nível dos nós obtiveram-se as medidas apresentadas na Tabela 4.2.

Intervalo	1-3	2-4	3-5	4-6	5-7	6-8	7-9	8-10	9-11	10-12
Grau Médio	20,887	12,960	17,642	15,008	13,383	15,987	15,106	16,549	15,571	12,797
Intermediação média	358,804	94,338	209,193	139,481	114,718	125,263	77,476	200,799	224,508	162,187
Centralidade *eigenvector* média	0,028	0,024	0,023	0,021	0,027	0,012	0,010	0,041	0,042	0,025

Tabela 4.2: Medidas ao nível dos nós

Pela análise do grau médio pode-se observar que o número médio de produtos diferentes comprados pelos clientes é relativamente baixo. O valor mais elevado para esta medida é obtido no primeiro intervalo (1-3) enquanto que o valor mais baixo é obtido no último período (10-12). Nesta medida não é possível observar qualquer tipo de padrão.

Uma análise mais detalhada revela que existe um cliente (P00003905) que em todos os instantes ocupa uma posição dominante com um grau bastante elevado (entre 75 e 137), ocupando um lugar de topo na maioria dos instantes e com intermediação a variar entre 6096,79 e 18655,95 em que apenas não está no topo no primeiro instante (1-3). Analisando a evolução temporal destas duas métricas, não é possível observar um padrão que permita identificar algum tipo de comportamento comercial ou comportamental por parte do cliente.

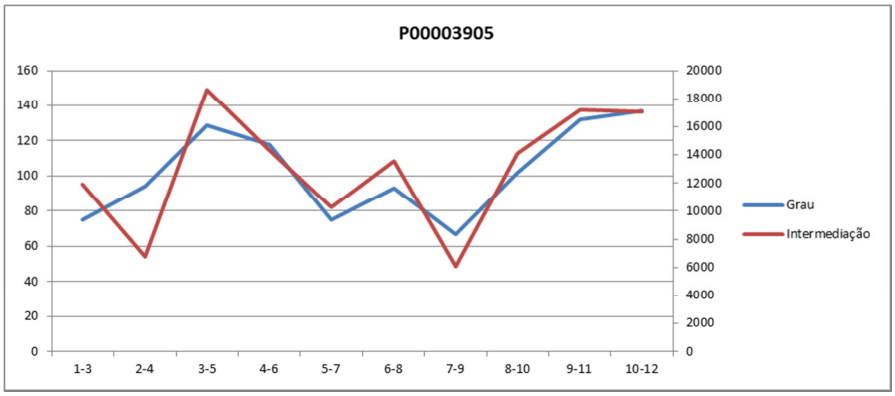

Figura 4.2: Grau e intermediação - P00003905

Mapeando os restantes clientes de acordo com o número de períodos em que demonstram actividade relevante (com valores elevados de grau e intermediação) podem-se identificar dois (visíveis em quatro períodos) e quatro (visíveis em três períodos). Os dois primeiros podem-se observar na Figura 4.3 e Figura 4.4.

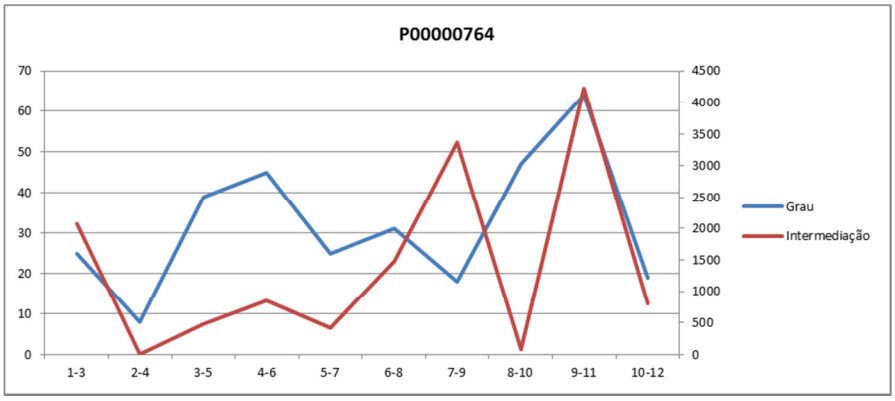

Figura 4.3: Grau e intermediação - P00000764

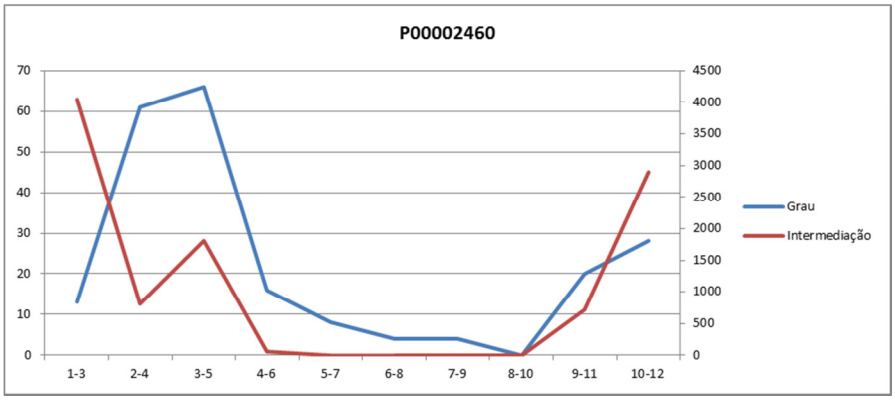

Figura 4.4: Grau e intermediação - P00002460

Adicionalmente são também representados três clientes importantes da empresa (para comparação futura).

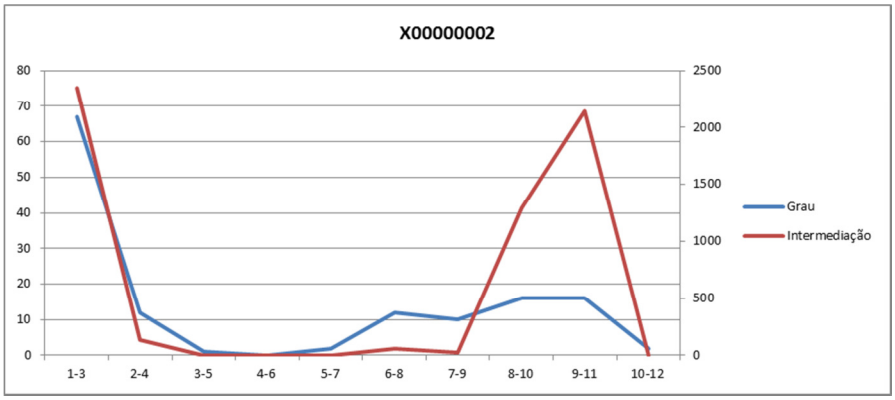

Figura 4.5: Grau e intermediação - X00000002

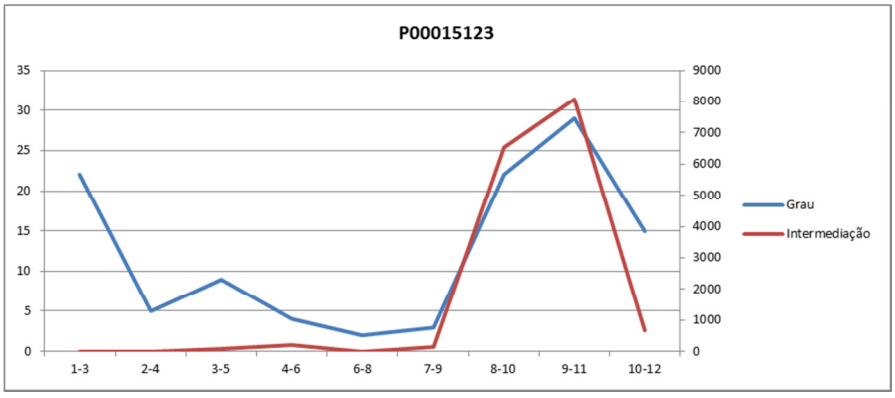

Figura 4.6: Grau e intermediação - P00015123

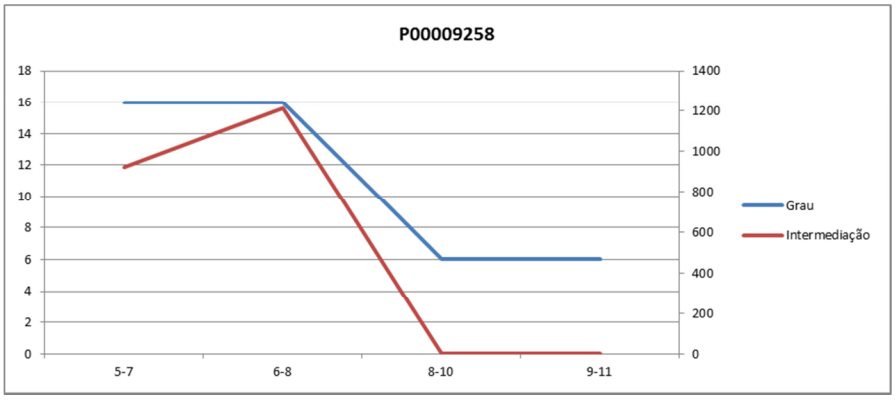

Figura 4.7: Grau e intermediação - P00009258

Em todos os clientes apresentados, pode-se observar que existem períodos em que apresenta valores de grau e intermediação bastante elevados contrastando com períodos em que estes valores chegam a assumir o valore de 0. Estes clientes são naturalmente alguns dos *hubs* identificados visualmente nas imagens da rede.

Uma vez que os períodos em que é visível algum tipo de acção dos clientes serem reduzidos, não faz sentido tentar identificar um padrão de comportamento.

No que diz respeito à centralidade *eigenvector*, pode-se observar na Figura 4.8 a representação desta métrica para os clientes que tiveram a classificação mais elevada em qualquer um dos instantes. De acordo com a interpretação desta métrica, estes clientes seriam os clientes ideais para promover um produto novo, uma vez que a sua influência e visibilidade da rede poderia influenciar outros clientes a comprar também o mesmo produto. Nesta representação consegue-se observar muito bem a baixa dinâmica desta rede de clientes com o exemplo do cliente P00021149 que no intervalo 5-7 assume um valor de 1 para a centralidade *eigenvector* e no intervalo 8-10, este valor desce para 0. A instabilidade da métrica de centralidade *eigenvector* significa que existe pouca actividade de alguns clientes resultando numa variação abrupta de ligações.

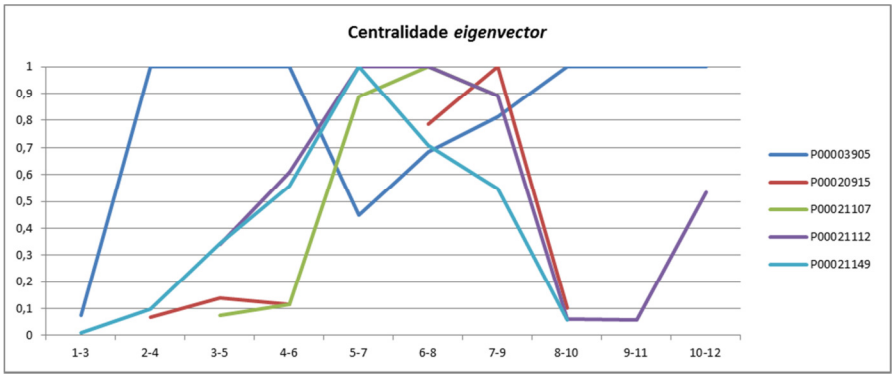

Figura 4.8: Centralidade *eigenvector*

4.2.3 Medidas ao nível da rede

Ao nível da rede obtiveram-se as medidas apresentadas na Tabela 4.3.

Intervalo	1-3	2-4	3-5	4-6	5-7	6-8	7-9	8-10	9-11	10-12
Diâmetro rede	12,00	9,00	10,00	8,00	7,00	6,00	7,00	11,00	10,00	7,00
Raio	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Comprimento médio de percursos	4,77	2,66	3,07	2,73	2,77	2,71	2,79	3,37	3,24	2,95
Nº percursos mais curtos	63962	34356	59808	41918	29484	43830	21300	48714	60818	52574
Densidade	0,06	0,04	0,06	0,06	0,06	0,05	0,06	0,06	0,05	0,04
Coeficiente de aglomeração médio	0,92	0,89	0,86	0,91	0,89	0,88	0,89	0,92	0,92	0,90
Modularidade	0,61	0,61	0,60	0,63	0,60	0,53	0,50	0,68	0,67	0,69

Tabela 4.3: Medidas ao nível da rede

Ao longo dos vários intervalos, pode-se observar que o diâmetro varia entre 12 no intervalo 1-3 e 6 no intervalo 6-8. Estas observações são em parte suportadas pela análise visual da rede em cada um destes intervalos. Podem-se tirar dois tipos de conclusões. A primeira é que no primeiro intervalo, a nível geral, a variedade de produtos comprados pelos clientes foi maior que no segundo. A segunda é que dada a diferença de ligações versus o número de nós (intervalo 1-3: 336 nós e 3509 ligações;

intervalo 6-8: 300 nós e 2398 ligações), um grupo homogéneo de clientes não teve qualquer tipo de actividade levando a uma diminuição considerável do número de ligações (32%) para uma diminuição mais modesta do número de nós (10%). Em termos de sazonalidade, pode-se observar um valor relativamente baixo nos intervalos 4-6, 5-7, 6-8, 7-9.

A medida da densidade da rede é bastante próxima de zero em todos os intervalos revelando que existe uma grande variedade nas escolhas dos clientes e que a rede de clientes é pouco madura.

O valor baixo observado para o comprimento médio de percursos demonstra que existem alguns *hubs* na rede ou clientes que consomem produtos comuns a vários outros clientes, facto que já foi demonstrado na análise independente dos clientes.

O coeficiente de aglomeração é bastante elevado pelo que demonstra que existe uma quantidade considerável de conjuntos de clientes que compram os mesmos produtos.

4.2.4 Detecção comunidades

A detecção de comunidades é efectuada recorrendo ao método de *Blondell* com a modularidade como critério de paragem. Os resultados visuais para os vários instantes são apresentados nas imagens seguintes. De notar que a cor das imagens é referente ao agrupamento detectado pelo algoritmo de *Blondell* e o tamanho das circunferências é proporcional ao valor da centralidade *eigenvector* do cliente.

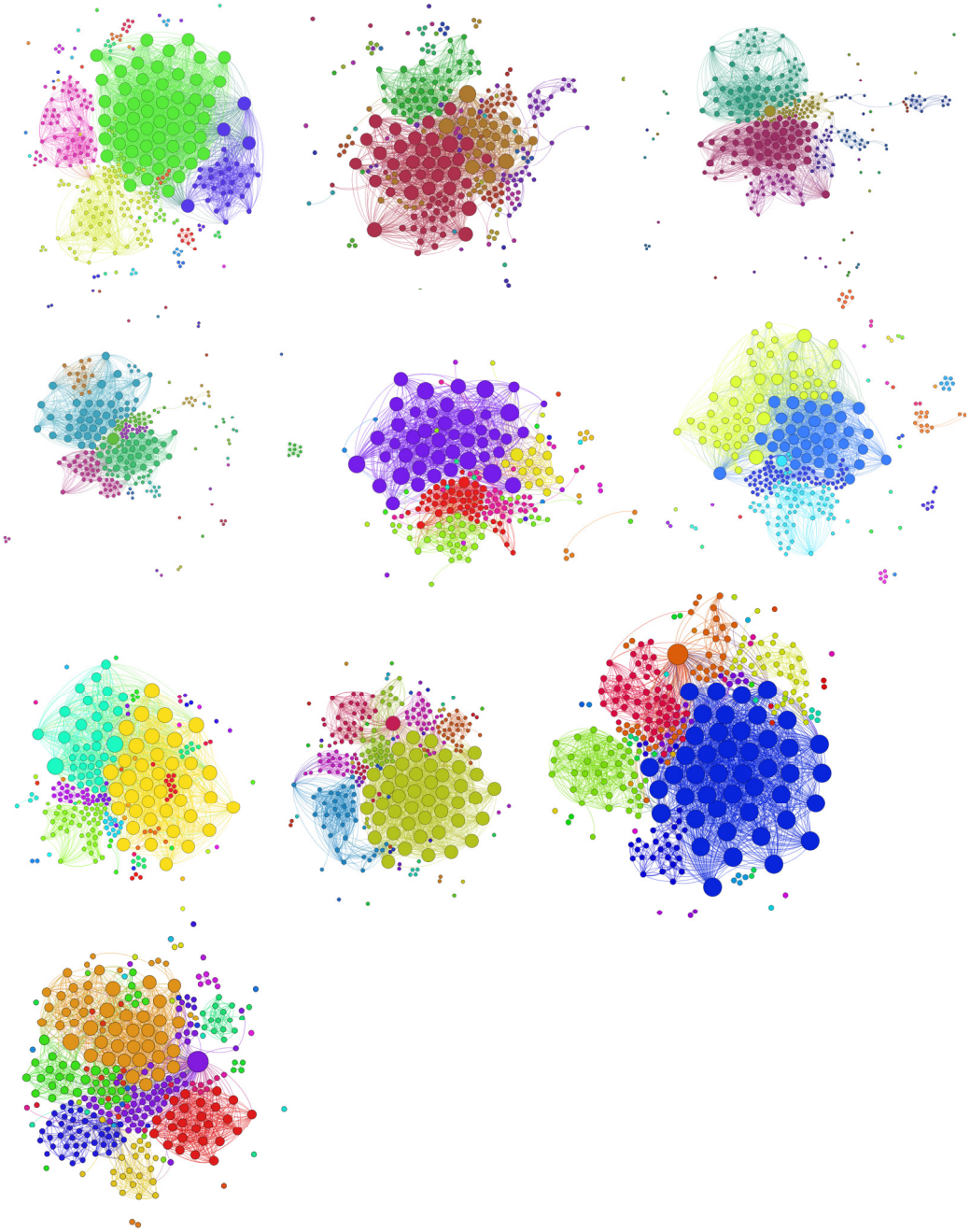

Figura 4.9: Representação da rede com detecção de comunidades com janela deslizante

Com a análise das várias imagens relativas aos vários intervalos "capturados" pelo método de janela deslizante é possível observar uma dinâmica bastante elevada no que concerne o aparecimento de novas comunidades, novas ligações entre comunidades ou perda de ligação entre comunidades. O primeiro agrupamento de comunidades de

acordo com o método de *Blondell* pode ser observado, para todos os instantes, na Figura 4.10.

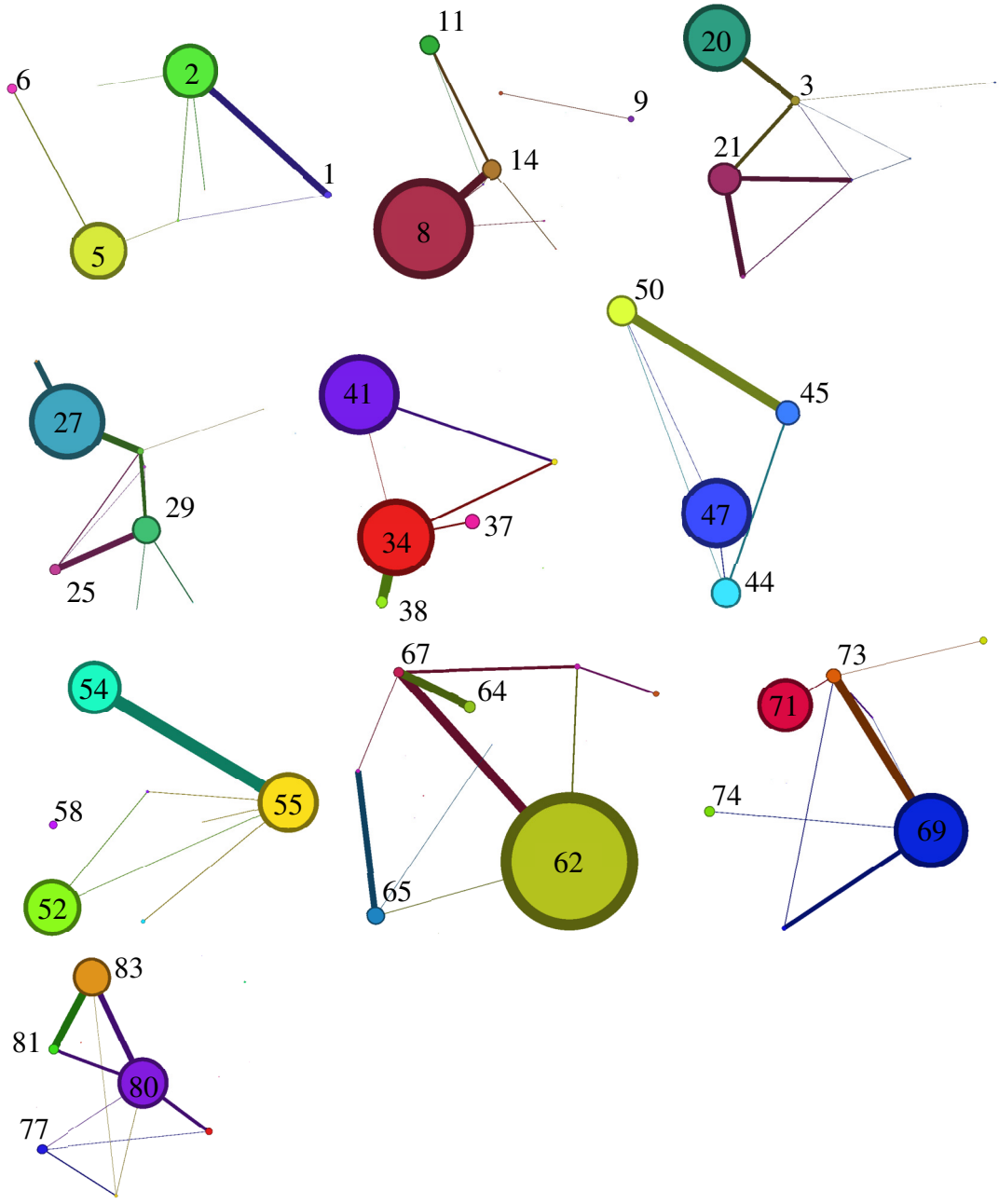

Figura 4.10: Representação da rede com primeiro agrupamento de comunidade com janela deslizante

Recorrendo a uma versão ligeiramente modificada do algoritmo MEC (Oliveira & Gama, 2010) foi feita a análise da evolução das comunidades. Dado o grande número de

comunidades em cada instante, foi escolhido um número suficientemente grande que permitisse representar pelo menos 75% da população total. Isto resultou numa média de oito comunidades por instante temporal. Adicionalmente, os parâmetros utilizados para definir separação e continuação foram 20% e 50% respectivamente. Os resultados obtidos podem ser observados na Figura 4.11. De notar que linhas a tracejado significam uma separação e linhas contínuas significam uma continuação. Após esta análise, os números das principais comunidades de clientes foram colocados na Figura 4.10.

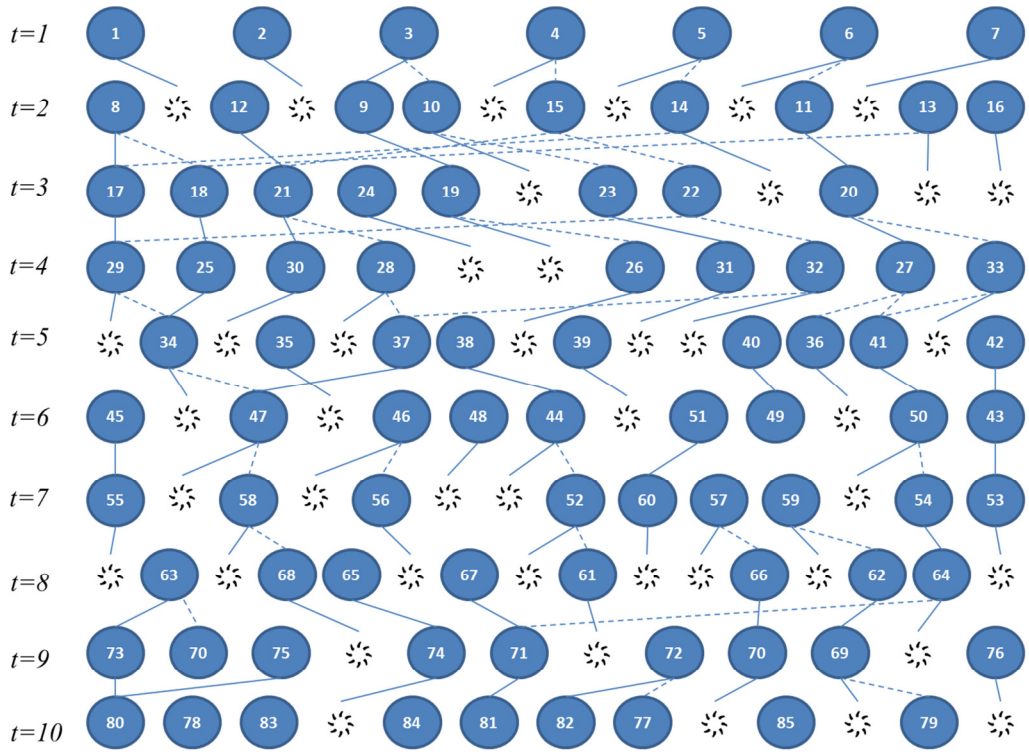

Figura 4.11: Evolução das comunidades com janela deslizante

A interpretação da Figura 4.11 é bastante difícil. A grande dinâmica observada aumenta a dificuldade de análise. No entanto, seria importante, tentar perceber a razão para o desaparecimento das comunidades em instantes consecutivos, principalmente em casos em que uma comunidade se mantém durante alguns instantes e depois desaparece (ex:

C42->C43->C53->* do instante t=5 a t=8). Os casos em que se verifica um desaparecimento e uma separação (ex: C69->*+->C79 no instante t=10) podem-se considerar como de relativo sucesso, isto é, para uma comunidade que praticamente desapareceu, foi possível conservar alguns clientes.

4.3 Análise com janela acumulada

4.3.1 Apresentação visual

Com a aplicação do método de janela acumulada obteve-se a seguinte representação da rede nos vários instantes.

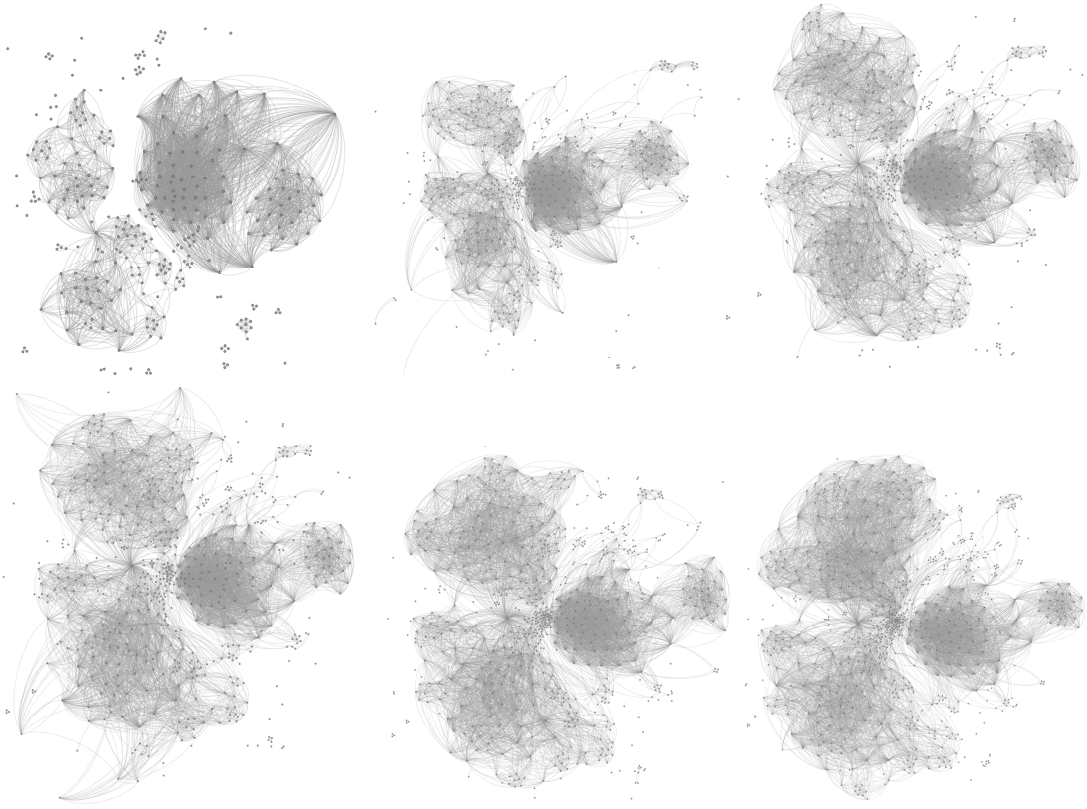

Figura 4.12: Janela acumulada. Instantes 1-3 a 6-8.

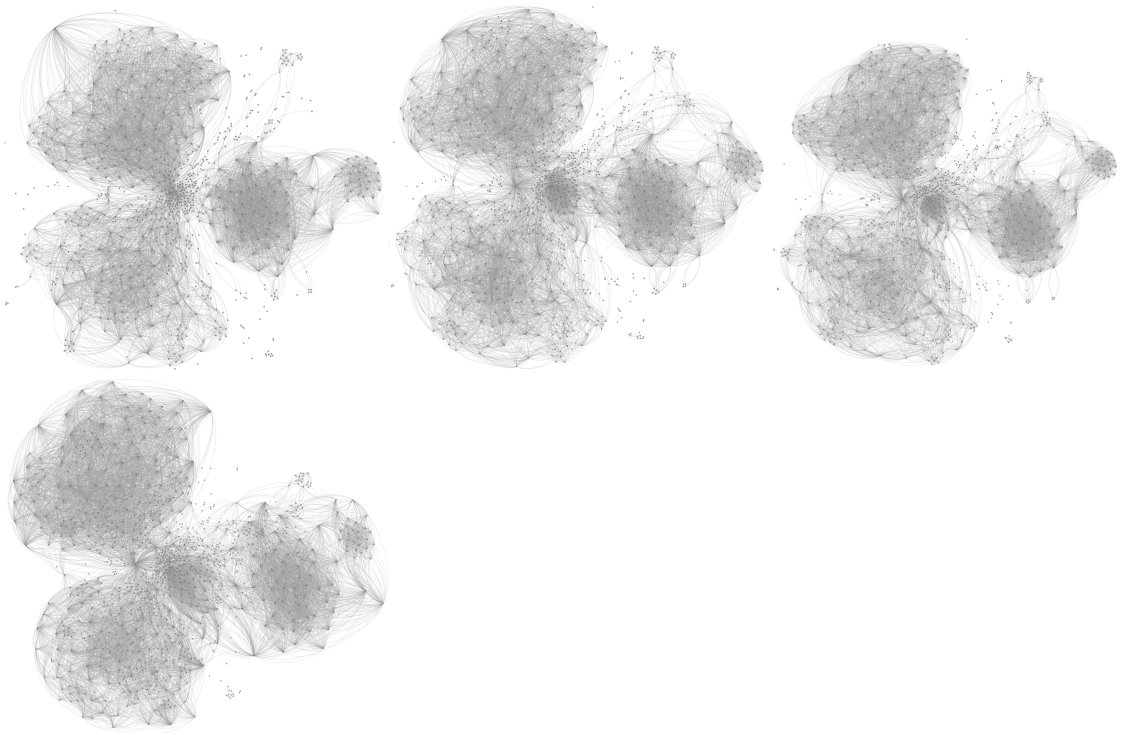

Figura 4.13: Janela acumulada. Instantes 7-9 a 10-12.

A análise visual da janela acumulada revela a composição de agrupamentos de clientes em determinadas zonas da rede sendo possível observar um *hub* central a partir do período 1-4.

Relativamente ao número de nós e ligações em cada período da janela, estes estão representados na Tabela 4.4.

Intervalo	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
Nós	336	454	543	581	665	746	787	894	947	1014
Ligações	3509	4778	6050	6460	7389	8568	8985	10801	11193	12259

Tabela 4.4: Número de nós e ligações

4.3.2 Medidas ao nível dos nós

Ao nível dos nós obtiveram-se as medidas apresentadas na Tabela 4.5.

Intervalo	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
Grau Médio	20,887	21,048	22,284	22,238	22,223	22,971	22,834	24,163	23,639	24,179
Intermediação média	358,804	343,786	390,118	449,057	550,555	604,916	636,989	663,513	699,973	751,927
Centralidade *eigenvector* média	0,028	0,057	0,095	0,109	0,150	0,114	0,102	0,069	0,068	0,062

Tabela 4.5: Medidas ao nível dos nós

A análise do grau médio revela que o número médio de produtos diferentes adquiridos, apesar de baixo, aumenta com o tempo. Esta é uma observação natural neste tipo de dinâmica temporal uma vez que não estão a ser eliminados da análise os intervalos mais antigos.

No que diz respeito à intermediação, este valor também verifica um aumento à medida que o tamanho da janela abrange mais intervalos.

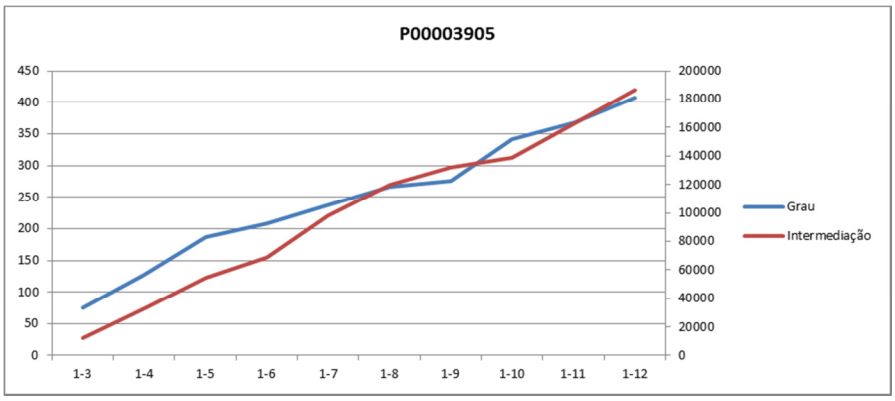

Figura 4.14: Top 5 de Grau e Intermediação - P00003905

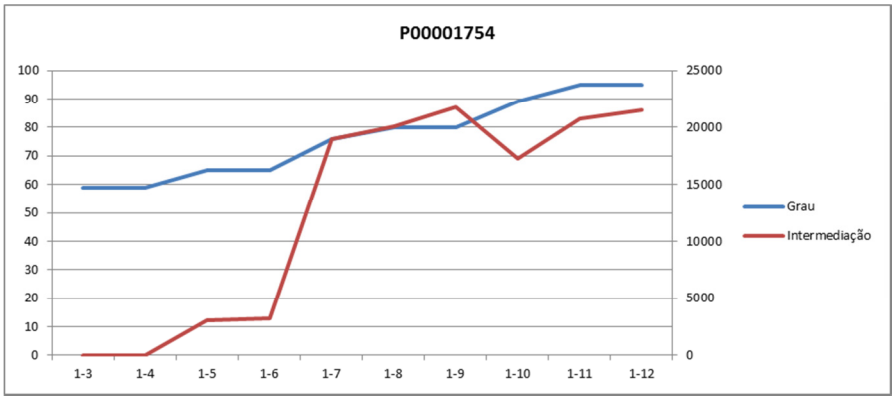

Figura 4.15: Top 5 de Grau e Intermediação - P00001754

Figura 4.16: Top 5 de Grau e Intermediação - P00004323

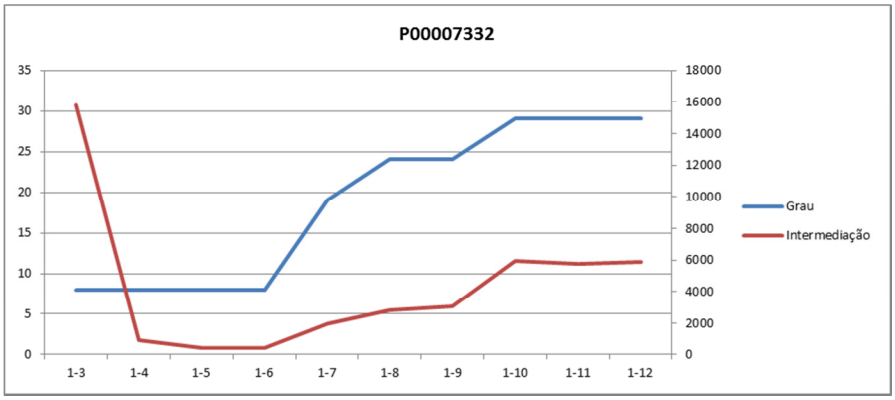

Figura 4.17: Top 5 de Grau e Intermediação - P00007332

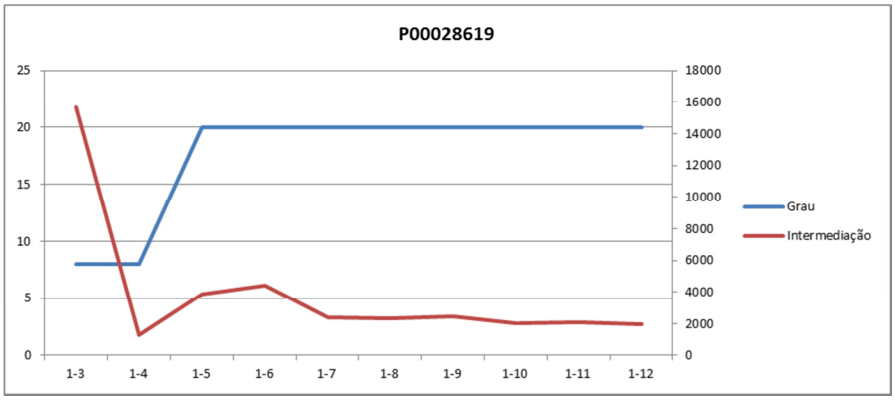

Figura 4.18: Top 5 de Grau e Intermediação - P00028619

Na Figura 2.4 pode-se observar o Top 5 de clientes tendo em conta as métricas de grau e intermediação. Neste formato é possível verificar a consistência do comportamento destes clientes ao longo do tempo. Pode-se observar que o cliente P00003905 tem um comportamento consistente de crescimento tanto a nível de grau como na intermediação. Outro cliente que também apresenta uma tendência crescente destas medidas, se bem que muito mais modesta é o P00001754. Os restantes clientes representados apresentam um valor ligeiramente crescente do grau mas irregular em termos de intermediação. Esta observação poderá indicar que estes clientes passaram a comprar produtos diferentes.

Na Figura 4.21 pode-se observar a medida do grau e intermediação para três cliente de relevo que não apareceram no Top 5.

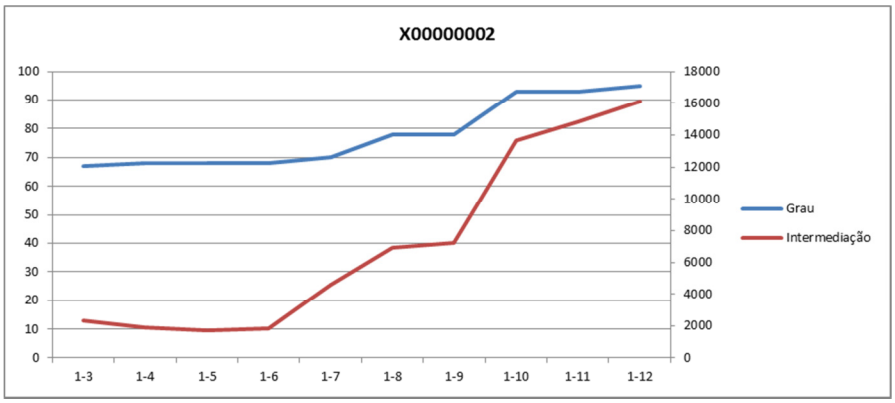

Figura 4.19: Grau e intermediação - X00000002

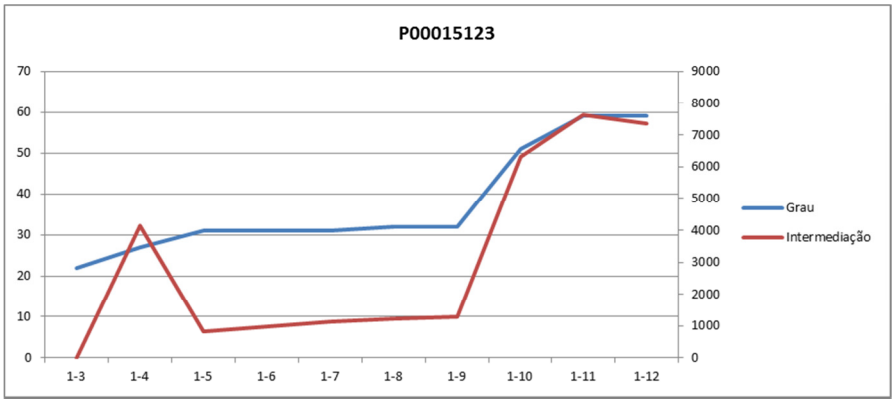

Figura 4.20: Grau e intermediação - P00015123

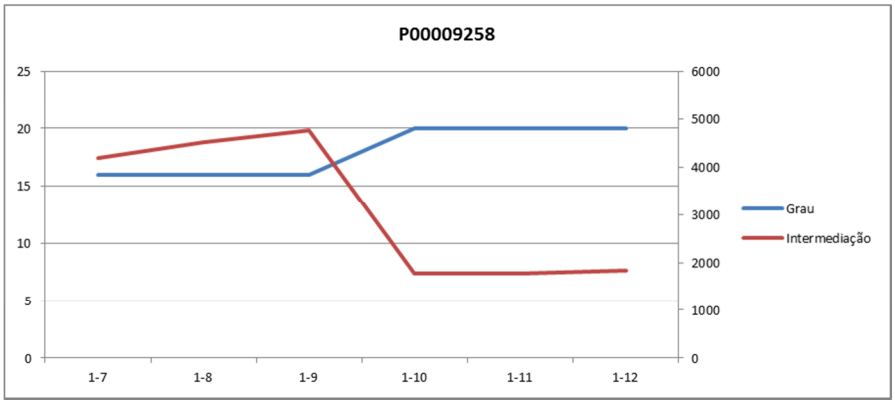

Figura 4.21: Grau e intermediação - P00009258

Em relação à centralidade *eigenvector*, pode-se observar na Figura 4.22 a representação desta métrica para os clientes que tiveram a classificação mais elevada em qualquer um dos instantes. Uma observação interessante é o facto de praticamente todos terem uma descida no valor desta métrica ao longo do tempo (com a excepção do P00003905). Esta observação faz sentido pois à medida que a rede vai crescendo, a importância relativa dos clientes vai diminuindo e perdem um pouco a sua "centralidade". Mesmo assim continuam a ser bons candidatos para a promoção de novos produtos.

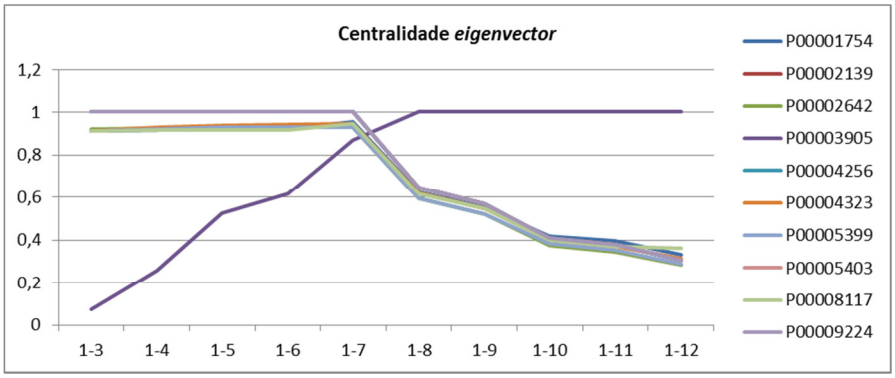

Figura 4.22: Centralidade *eigenvector*

4.3.3 Medidas ao nível da rede

Ao nível da rede obtiveram-se as medidas apresentadas na Tabela 4.6.

Intervalo	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
Diâmetro rede	12,00	9,00	8,00	8,00	8,00	8,00	8,00	8,00	8,00	8,00
Raio	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Comprimento médio de percursos	4,77	3,28	3,09	3,12	3,17	3,15	3,16	2,99	2,97	2,96
Nº percursos mais curtos	63962	136658	203092	246624	337062	420642	464530	595356	671724	778948
Densidade	0,06	0,05	0,04	0,04	0,03	0,03	0,03	0,03	0,03	0,02
Coeficiente de aglomeração médio	0,92	0,89	0,86	0,85	0,83	0,83	0,82	0,82	0,82	0,81
Modularidade	0,61	0,67	0,69	0,68	0,69	0,68	0,68	0,68	0,68	0,67

Tabela 4.6: Medidas ao nível da rede

Analisando a medida do diâmetro pode-se concluir que para além dos dois primeiros períodos em que este assume os valores de 12 e 9 respectivamente, este não muda para além do período 1-5. Isto significa que apesar do número de clientes da rede subir de 543 para 1014, a variedade de produtos comprados e que não é partilhada por determinados grupos de clientes se mantém ao longo do tempo.

A medida da densidade da rede mantém-se bastante próxima de zero e decresce ao logo de todos os intervalos o que, dado o crescimento constante de clientes e ligações entre

os mesmos, demonstra que existem grupos isolados em que os clientes que fazem parte destes grupos não estão ligados aos restantes.

O valor do comprimento médio dos percursos suporta a assunção que existe um número elevado de grupos uma vez que este se mantém bastante baixo e decrescente ao longo do tempo.

Quanto ao coeficiente de aglomeração, pode-se observar um ligeiro decréscimo ao logo do tempo. Esta observação ajuda a justificar a existência de grupos mais isolados, no entanto, isto não significa que dentro desses grupos, o nível de formações triádicas não se mantenha elevado.

4.3.4 Detecção comunidades

A detecção de comunidades é feita da mesma forma que na análise temporal com janela deslizante. Sendo assim a cor das imagens é referente ao agrupamento detectado pelo algoritmo de *Blondell* e o tamanho das circunferências é proporcional ao valor da centralidade *eigenvector* do cliente.

Ao analisar a Figura 4.10, pode-se observar a formação e evolução das várias comunidades de clientes ao longo do tempo. São claramente visíveis três a quatro grandes grupos intermediados por alguns *hubs*. Estes *hubs* são clientes pluri-produto, ou seja, clientes que compram uma variedade grande de produtos.

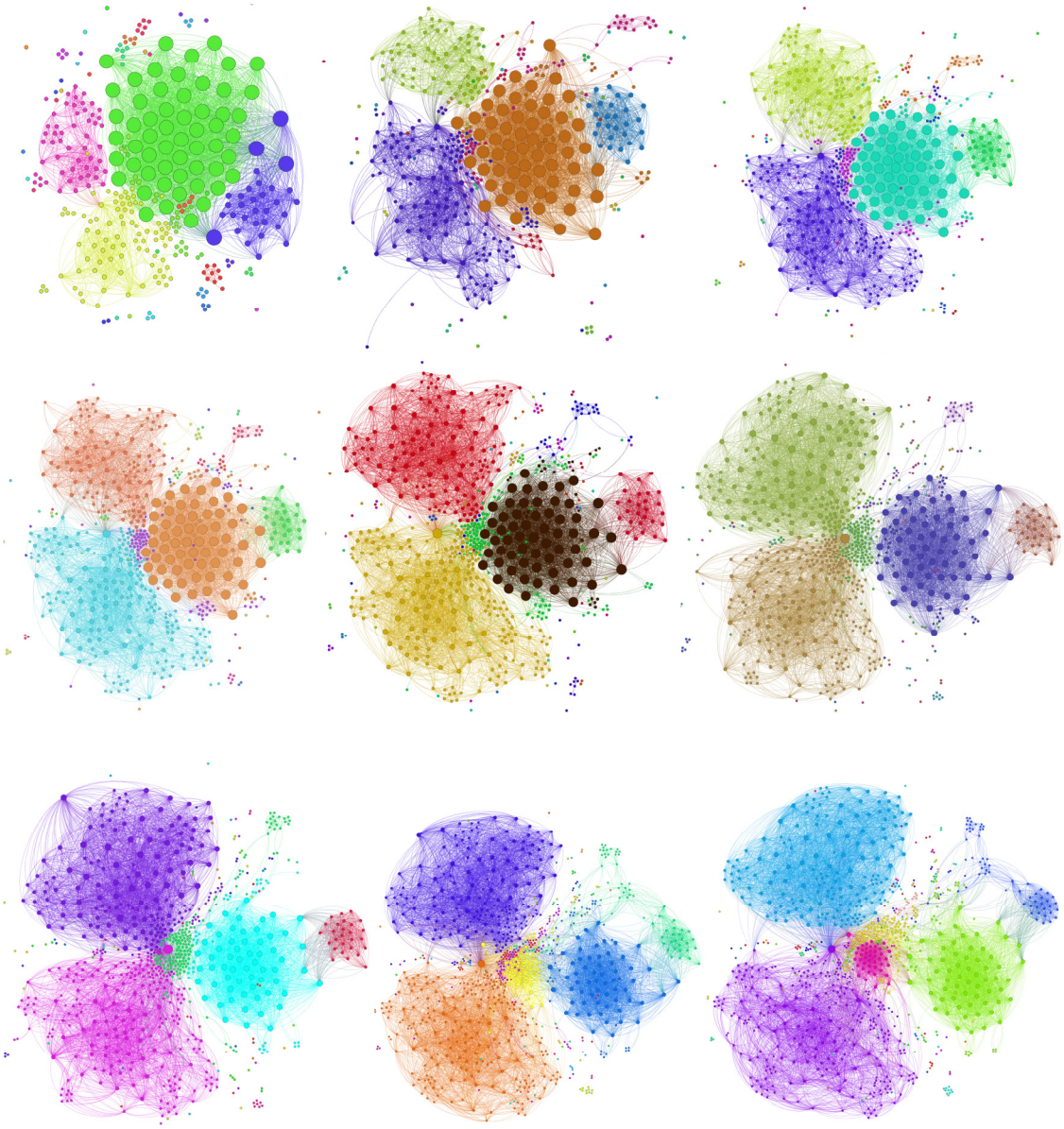

Figura 4.23: Representação da rede com detecção de comunidades com janela acumulada – instantes 1-3 a 1-11

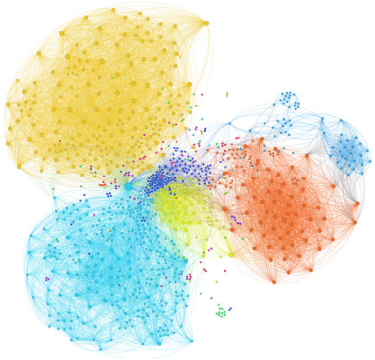

Figura 4.24: Representação da rede com detecção de comunidades com janela acumulada – instante 1-12

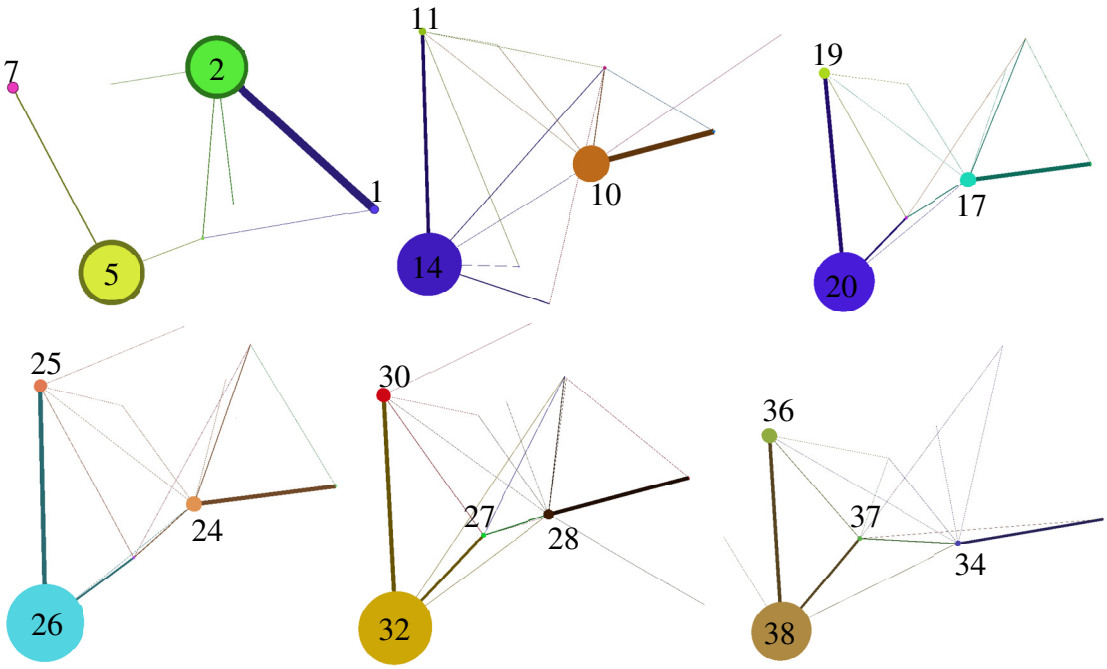

Figura 4.25: Representação da rede com primeiro agrupamento de comunidade com janela acumulada – instantes 1-3 a 1-8

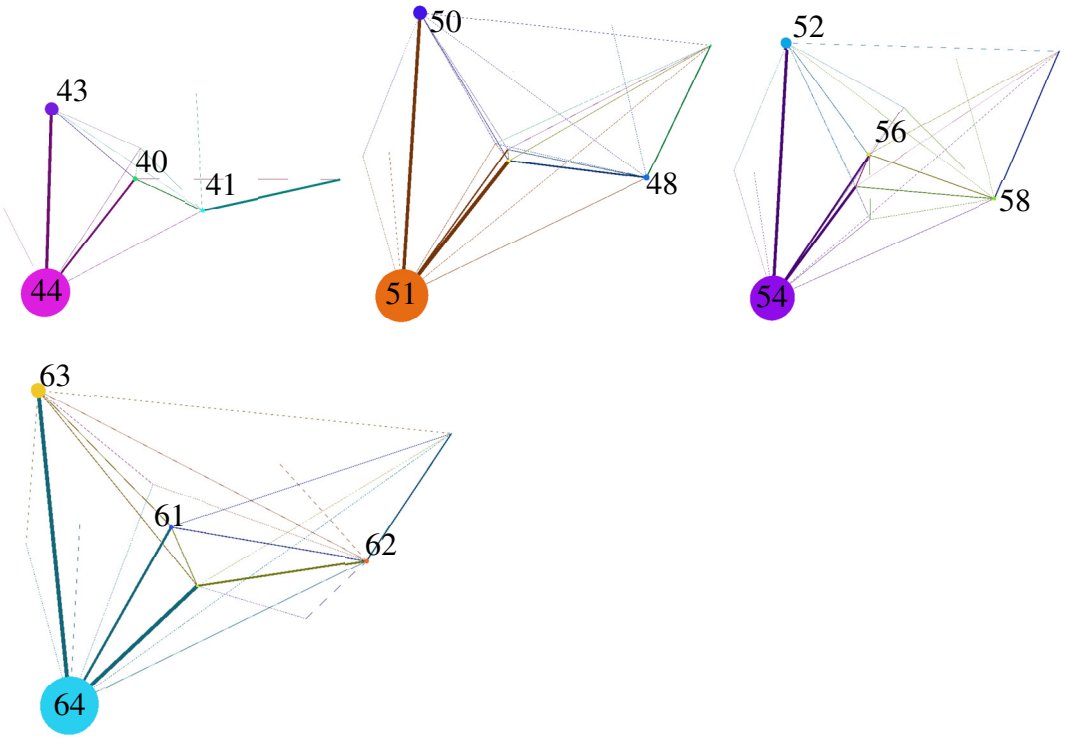

Figura 4.26: Representação da rede com primeiro agrupamento de comunidade com janela acumulada – instantes 1-9 a 1-12

A estrutura da rede e a sua evolução é bastante mais fácil de identificar analisando o agrupamento em comunidades. Ao longo do tempo, pode-se observar o nascimento de novas comunidades bem como a separação de algumas comunidades noutras mais pequenas. Estas separações são devidas ao aparecimento de novos clientes, mais parecidos com alguns que já estavam numa comunidade e cuja ligação é mais forte do que a anterior.

Usando o mesmo algoritmo que na dinâmica de janela deslizante para a análise de evolução das comunidades, obteve-se o resultado mapeado na Figura 4.27.

Recorde-se que foram apenas analisadas as comunidades que representassem a rede em pelo menos 75% em cada instante e que os parâmetros para definir separação e continuação foram 20% e 50% respectivamente. Aqui também as linhas a tracejado significam uma separação e linhas contínuas significam uma continuação.

A primeira observação que se pode fazer pela análise da figura é que a evolução das comunidades é muito mais estável com a dinâmica de janela acumulada. Nesta visão seria importante caracterizar as comunidades mais estáveis para tentar perceber se seria possível "cativar" mais clientes. A interpretação da dinâmica de separação de alguns clientes e a identificação destes clientes específicos poderia também levar à interpretação do comportamento que os levou a mudar de comunidade.

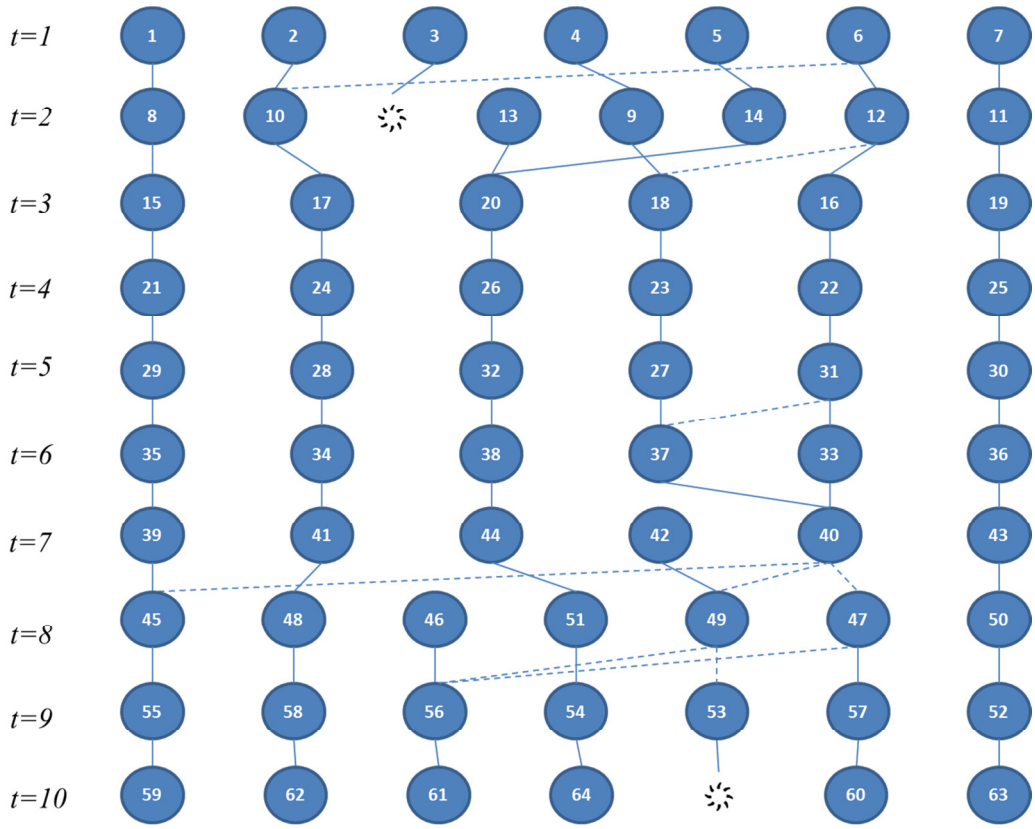

Figura 4.27: Evolução das comunidades com janela acumulada

Nesta dinâmica temporal, o desaparecimento de comunidades assume uma grande importância porque como se trata de um "acumulado" o número potencial de clientes que está numa comunidade é bastante maior do que na dinâmica de janela deslizante. A interpretação destas comunidades e das razões que levaram ao seu desaparecimento é bastante importante. De notar que a utilização do termo "desaparecimento" não é bem

exacta uma vez que os clientes apenas deixaram de pertencer às comunidades representativas dos 75% da rede.

4.4 Análise comparativa dos resultados

O objectivo desta secção é de fazer uma análise comparativa dos resultados obtidos em cada um dos métodos de análise temporal.

Pela análise visual da disposição da rede pode-se concluir que a rede não é suficientemente dinâmica para a observação em janela deslizante ou, pelo menos com os parâmetros escolhidos. Não foi possível fazer uma ampliação dos parâmetros da janela por falta de informação para outros períodos. Consegue-se observar, ao longo do tempo a formação de agrupamentos bem definidos e a existência de alguns *hubs* que fazem de interface entre estes agrupamentos.

No que concerne o número de nós e ligações, na aplicação da janela deslizante o número de nós varia entra 246 e 336 e as ligações entre 1519 e 3509. Estes números revelam que existem períodos em que aparecem e desaparecem grupos bem ligados de clientes uma vez que para uma variação baixa do número de nós existe uma variação bastante maior do número de ligações. Pela observação da janela acumulada, a rede final apresenta 1014 clientes protagonizando 12259 ligações.

As métricas ao nível dos nós revelam que o número médio de produtos comprados pelos clientes é relativamente baixo variando entre 12 e 20 na óptica da janela deslizante, e mesmo com o passar do tempo este valor não aumenta muito sendo de 24 no último instante da janela acumulada.

A medida da intermediação revela que existem alguns *hubs* na rede tanto em intervalos curtos (na janela deslizante) como se mantém ao longo do tempo (na janela acumulada).

A centralidade *eigenvector* permite tirar uma conclusão interessante para cada um dos métodos. Os clientes com posições centrais em cada instante variam bastante com o método de janela deslizante enquanto no método de janela acumulada, além de apresentarem um comportamento mais suave e com a excepção de um (P00003905),

vêem o seu valor de centralidade *eigenvector* diminuir ao longo do tempo. Naturalmente o crescimento da rede faz diminuir a importância relativa de alguns clientes.

A partir das medidas ao nível da rede, pela análise do diâmetro na óptica da janela deslizante pode-se concluir que em termos gerais a variedade de produtos comprados pelos clientes varia ao longo do tempo. Simultaneamente pode-se observar a inactividade de grupos homogéneos de clientes em intervalos distintos. Os valores obtidos para os intervalos 4-6, 5-7, 6-8, 7-9 sugerem algum tipo de sazonalidade.

O valor da densidade é bastante baixo nos dois métodos e diminui ao longo do tempo na janela acumulada. Esta observação revela que existe uma grande variedade nas escolhas dos clientes e que a rede de clientes é pouco madura.

O valor baixo observado para o comprimento médio de percursos demonstra que existem alguns *hubs* na rede ou clientes que consomem produtos comuns a vários outros clientes.

A medida elevada do coeficiente de aglomeração, apesar de diminuir ligeiramente na janela acumulada, permite concluir que existe uma quantidade elevada de clientes que compram os mesmos produtos.

A análise detalhada de alguns clientes permite concluir que os clientes com maior actividade são P00003905, P00000764 e P00002460 enquanto que os que apresentam um comportamento mais consistente são P00003905, P00001754.

Os clientes P00004323 e P00007332 provavelmente passaram a comprar produtos diferentes ao longo do tempo.

Dos três clientes referidos anteriormente como sendo clientes que também são importantes (X00000002, P00015123 e P00009258) mas cujos resultados não os classificavam como favoritos na óptica da janela deslizante, revelam, de facto uma posição diferente na janela acumulada.

A detecção de comunidades com o método de janela deslizante permite identificar que em determinados intervalos desaparecem comunidades inteiras voltando a aparecer mais

tarde, significando que existem períodos em que não há qualquer actividade por parte destes clientes. A análise de janela acumulada revela a existência de grupos bem definidos permitindo fazer uma caracterização credível dos clientes novos que são colocados nestes grupos.

4.5 Sumário das duas abordagens

As duas abordagens temporais permitem fazer análises bastante distintas, Enquanto uma valoriza o factor da "actividade recente" (janela deslizante), o outro considera todo o histórico temporal. Apesar deste facto não se pode considerar que uma é melhor do que a outra.

Da análise de janela temporal consegue-se ter uma perspectiva de uma rede bastante dinâmica em que a maioria dos clientes, com a excepção de uma pequena amostra, não tem uma actividade constante. Uma análise mais detalhada faria sentido de duas formas. A primeira por tentar analisar o sucesso dos clientes que se mantém activos e a segunda ao tentar analisar os que tem uma actividade baixa ou inconstante.

Analisando os resultados obtidos com o método de janela acumulada, é possível descobrir em que grupos novos clientes ficam colocados. Do ponto de vista comercial esta é uma observação bastante interessante uma vez que poderão ser usadas técnicas de marketing com sucesso comprovado no grupo nestes novos clientes.

Capítulo 5

5 Conclusões

Este capítulo dividido em duas partes. Na primeira é feito o resumo do trabalho efectuado bem como a identificação das principais contribuições para a área de análise de dados. Na segunda são feitas sugestões para um trabalho futuro.

5.1 Lições aprendidas

Nesta tese foi proposto um método de análise de redes de dados de clientes tendo como base para a construção da rede duas premissas: o cliente é um nó e a ligação entre dois clientes é um produto adquirido pelos dois.

O método consiste na análise de métricas usadas em redes sociais em duas dinâmicas temporais distintas: janela deslizante e janela acumulada.

O conjunto de métricas usadas permitiu fazer a análise individual das características dos clientes bem como inferir alguns aspectos relacionados com o seu comportamento. Para além de características dos clientes também foi possível tirar conclusões relativas à rede numa forma global.

As dinâmicas temporais em janela acumulada e deslizante, aliadas à detecção de comunidades em cada instante permitiram ter uma visualização bastante interessante da evolução das comunidades de clientes ao longo do tempo. Numa das dinâmicas (janela deslizante) foi possível observar transições mais "bruscas" entre intervalos enquanto que na outra (janela acumulada) estas transições foram muito menores.

A premissa inicial para a análise temporal em duas formas distintas tinha como objectivo tentar perceber qual das duas seria a melhor, no entanto, à medida que o trabalho foi evoluindo, não foi possível chegar a esta conclusão. As duas dinâmicas,

efectivamente completam-se. Enquanto uma permite ter uma visão do passado mais recente, a outra considera todo o histórico e, independentemente do tipo de análise pretendida, o conhecimento extraído por um dos métodos completa o adquirido pelo outro.

5.2 Trabalho Futuro

Como propostas de trabalho futuro, consideram-se duas possibilidades.

A primeira consiste na comprovação do método, isto é, compreende uma aplicação da metodologia a outras redes de clientes, de empresas reais e com o objectivo de avaliar o sucesso dos resultados obtidos em termos de retorno para a empresa.

A segunda, assente em algoritmos de previsão, baseados em regras de decisão ou árvores de decisão consiste em complementar a metodologia com a possibilidade de prever o comportamento futuro de clientes ou a análise de novos clientes tendo em conta o histórico dos clientes existentes.

Referências Bibliográficas

[1] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[2] Tutte, W.T. (2001), Graph Theory, Cambridge University Press, p. 30, ISBN 978-0-521-79489-3

[3] Oliveira, M., J. Gama (2012). An Overview of Social Networks Analysis, WIREs Data Mining Knowl Discov., pp. 99-115.

[4] Easley, D. & J. Kleinberg (2010). Networks, Crowds and Markets, New York: Editions Cambridge, pp. 21-74.

[5] Euler, L. (1766). Solutio problematis ad geometrian situs pertinentis, Commentarii academiae scientiarum Petropolitanae 8, 1741, pp. 128-140.

[6] John Joseph Sylvester (1878), Chemistry and Algebra. Nature, volume 17, page 284.

[7] Berry, Michael; Linoff, Gordon (1997). Data Mining Techniques For Marketing, Sales, and Customer Support, Wiley, p 234.

[8] Newman MEJ (2003). The structure and function of complex networks. SIAM Rev, 45:167–228.

[9] Freeman LC (1979). Centrality in social networks: conceptual clarification. Soc Netw, 1:215–239.

[10] Watts DJ, Strogatz SH (1998). Collective dynamics of smallworld networks. Nature, 393:440–442.

[11] Albert, R. & A. Barabási (2002). Statistical mechanics of complex networks, Reviews of Modern Physics, Vol.74, No. 1, pp. 47- 97.

[12] Maslov, S., Sneppen, K., and Zaliznyak, A. (2002). Pattern detection in complex networks: Correlation profile of the Internet.

[13] Kleinberg, J. M. (2000), Navigation in a small world, Nature 406, 845.

[14] Kathleen M. Carley, (2003), "Dynamic Network Analysis" in Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, Ronald Breiger, Kathleen Carley, and Philippa Pattison, (Eds.) Committee on Human Factors, National Research Council, National Research Council. Pp. 133-145.

[15] David Krackhardt & Kathleen M. Carley, (1998), "A PCANS Model of Structure in Organization" Pp. 113-119 in Proceedings of the 1998 International Symposium on Command and Control Research and Technology. Conference held in June. Monterray, CA.Evidence Based Research, Vienna, VA.

[16] Carley, Kathleen M. (2002), "Smart Agents and Organizations of the Future" The Handbook of New Media. Edited by Leah Lievrouw & Sonia Livingstone, Ch. 12 pp 206-220, Thousand Oaks, CA, Sage.

[17] Newman, M. E. J. & M. Girvan (2004). Finding and evaluating community structure in networks, Physical Review E 69(2): 26113.

[18] M. E. J. Newman (2003), Mixing patterns in networks. Phys. Rev. E 67, 026126.

[19] Yu-Ru Lin, Yun Chi, Shenghuo Zhu, Hari Sundaram, Belle L. Tseng (2009). Analyzing Communities and Their Evolutions in Dynamic Social Networks, ACM Transactions on Knowledge Discovery from Data, Vol. 3, No. 2, Article 8.

[20] Asur, S., Parthasarathy, S., Ucar, D. (2007). An event-based framework for characterizing the evolutionary behavior of interaction graphs. In Proceedings of the 13th ACM International SIGKDD Conference on Knowledge Discovery and Data Mining.

[21] Chi, Y., Song, X., Zhou, D., HINO, K., Tseng, B. L. (2007). Evolutionary spectral clustering by incorporating temporal smoothness. In Proceedings of the 13th ACM International SIGKDD Conference on Knowledge Discovery and Data Mining.

[22] Oliveira, M., J. Gama (2010). MEC - Monitoring Clusters' Transitions. STAIRS 2010, p 212-224.